Desarrollo de una herramienta de asistencia para el análisis de pruebas psicométricas de una población grande utilizando técnicas de *Big Data*.

SEBASTIAN ARDILA AGUDELO

Proyecto de grado presentado como requisito parcial
para aspirar al título de Ingeniero de Sistemas y computación

Director
Ing Germán A. Holguín L, M.Sc, Ph.D(C)

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PEREIRA
2017

Nota de Aceptación

Firma del Presidente del jurado

Firma del jurado 1 - Evaluador

Firma del jurado 2 - Evaluador

Firma del jurado 3 - Director

Pereira, 28 de Octubre de 2017

Dedicado a mi familia, por el apoyo incondicional que siempre me dio en todo ámbito profesional, a mis padres por el amor y por la disciplina que me enseñaron.

A las nuevas generaciones de ingenieros y amantes del conocimiento, y a toda aquella persona que lea este documento.

Agradecimientos a toda aquella persona que apoyo de una u otra forma el desarrollo de este proyecto de grado, a mi director de proyecto el Ingeniero Germán Andrés Holguín Londoño por su colaboración y gran apoyo en el desarrollo de este proyecto, a el psicólogo Jorge Andrés Orrego Martinez por proporcionar apoyo, e información sobre pruebas de análisis psicológicos estandarizadas necesarias para el desarrollo de este proyecto, a Juan Felipe Grajales por su ayuda y cooperación, a Alixon Franco Salazar por ayudarme en la redacción de este documento y a el programa de ingeniería de sistemas y computación por su gran formación durante todo el proceso académico.

CONTENIDO

pág.

LISTA DE TABLAS

LISTA DE FIGURAS

1. INTRODUCCIÓN

La psicometría es una ciencia que estudia las técnicas de medición en la psicología, preocupándose por medir correctamente la psiquis, sacando diagnósticos de las poblaciones. Existen pruebas psicométricas para medir diferentes aspectos como habilidades, actitudes, rasgos de la personalidad, entre otros [1].

Las sociedades han necesitado por décadas de los análisis psicométricos para poder analizar los comportamientos, tanto de individuos como grupales, con el fin de clasificar, intervenir o diagnosticar a las personas. La psicometría ha sido utilizada en diversos campos del conocimiento, no sólo en psicología, sino también en medicina general [2], enseñanza [3], e ingeniería [4] entre otros.

En este proyecto se realizó una investigación para desarrollar una metodología de análisis psicométricos, partiendo de bases de datos tan grandes que son imposibles de analizar por psicólogos en un tiempo eficiente, o por mediciones tradicionales en psicología.

En la actualidad existen técnicas de aprendizaje de máquina que se utilizan para generalizar comportamientos a partir de una información no estructurada. El aprendizaje de máquina es una ciencia en la que se desarrollan técnicas de inteligencia artificial y que le permiten a las computadoras reconocer patrones aprendidos. Entre las técnicas que podemos encontrar en el aprendizaje de máquina están las técnicas de regresión, de aprendizaje bayesiano, uso de métodos no paramétricos, máquinas de soporte vectorial, redes neuronales, árboles de decisión, entre otros [5].

Los algoritmos de aprendizaje de máquina sirven para resolver problemas de minería de datos. Por ejemplo, son útiles en situaciones donde los programas deben adaptarse dinámicamente a los cambios del entorno [6]. Bajo este contexto, el aprendizaje de máquina tiene distintas aplicaciones. Por ejemplo, los análisis y diagnósticos médicos o técnicos, clasificación de distintos entes, reconocimiento facial [7], reconocimiento de voz [8], escritura [9], etc., sin embargo, estas técnicas funcionan bajo la premisa básica de que toda la información disponible para el entrenamiento o minado de una base de datos, son de tamaño adecuado y caben en la memoria

de trabajo del equipo donde se pretende procesar [10].

1.1. DEFINICIÓN DEL PROBLEMA

Actualmente, existen muy pocos desarrollos tecnológicos implementados con grandes bases de datos en el campo de la psicometría. El *big data* y la psicología son usados por políticos o grandes compañías como Google o Facebook [11]. Esto nos muestra la importancia de esta ciencia en el siglo 21. Por este motivo, se deben implementar soluciones que sean eficientes, eficaces y escalables. Desarrollar metodologías que solucionen este tipo de problemas se ha convertido en una necesidad [12]. Las metodologías que están implementadas en psicometría actualmente en nuestra región no analizan los datos de manera masiva, pues en Risaralda, esta ciencia apenas se está dando a conocer [13].

Las instituciones han tenido tradicionalmente procesos socio-humanísticos que hoy en día requieren sistematización, puesto que analizar los datos se ha vuelto una tarea bastante tediosa. La cantidad de datos en la era del Internet de las cosas hace indispensable el uso de técnicas de *big data* [14]. El análisis manual de los datos hoy en día, requiere de una máquina que procese dicha información, debido a que se incrementan rápidamente. Por ejemplo, los datos de un colegio o una entidad podían caber en una libreta o en una pequeña base de datos que no requiriera mucha capacidad de memoria, entonces el análisis, y las búsquedas que se requerían no necesitaban ser procesadas en poco tiempo, ya que el tamaño de la base de datos hacía que estas tareas no tardaran en procesarse.

Debe tenerse en cuenta que la escala del problema hace que la técnica requerida para su solución sea muy diferente. Un ejemplo es tratar de procesar los datos de los estudiantes de la UNAL (53582 estudiantes a 2016) [15] con técnicas y algoritmos diseñados para un colegio de 800 estudiantes. Otro ejemplo, es tratar de encontrar patrones médicos entre los pacientes del puesto de salud de un pueblo como Belén de Umbría (27727 habitantes a 2017) [16] o tratar de encontrar patrones entre los afiliados a nueva EPS en Colombia (3'945.536 afiliados a 2016) [17].

En la actualidad, el tamaño de dichas bases de datos es tan grande, que requieren de máquinas muy potentes para ser procesadas, por lo que el análisis de estos datos se vuelve imposible de realizar manualmente.

Este trabajo pretende generar una herramienta que asista a los psicómetras en el proceso de valorar una población muy grande, que genera un volumen de información tal que las técnicas tradicionales no pueden manejar en un intervalo de tiempo razonable, requiriendo de las modernas técnicas de *big data*.

1.2. JUSTIFICACIÓN

Desarrollar una metodología que clasifique los datos recogidos de test psicométricos estandarizados, con el propósito de ser analizados en el contexto del *big data*, que sirva de soporte a los psicólogos de diferentes áreas, les ayudaría a tomar decisiones, hacer estudios o investigaciones en grandes poblaciones. Se plantea entonces, construir una metodología que sirva de herramienta de análisis poblacional, puesto que donde se quiera capturar datos de manera exponencial, se requiere de técnicas complejas de minado de grandes cantidades de datos, ya que ciertas cantidades de información son tan grandes, que la aplicación de técnicas tradicionales de minería de datos resultarían deficientes o muy poco prácticas.

El desarrollo de esta metodología, no solo serviría como soporte para los psicólogos, si no que en últimas, la sociedad sería beneficiada. Por un lado, los psicólogos podrían determinar el comportamiento de poblaciones gigantescas según las pruebas estandarizadas. Por otro lado, dichas poblaciones se verían beneficiadas, puesto que el tiempo de análisis y las respuestas de estos se optimizarían al computar una gran cantidad de datos de manera relativamente rápida y eficiente. Por último, serviría también de ejemplo para los investigadores y en conjunto la comunidad académica en cuanto al tema de *big data*.

1.3. OBJETIVOS

1.3.1. Objetivo General

Desarrollar una metodología para el análisis de test psicométricos aplicados masivamente, utilizando técnicas de big-data mining.

1.3.2. Objetivos Específicos

- Desarrollar una metodología para la clasificación de test psicométricos.
- Desarrollar una metodología para el minado de bases de datos masivas en psicometría.
- Desarrollar una metodología para el minado de bases de datos masivas que pueda ser aplicada a otras áreas de la sociedad

2. ESTADO DEL ARTE

A lo largo de los años, en el campo de la psicología se ha tomado la necesidad de analizar el comportamiento humano de grandes poblaciones. Las técnicas tradicionales de analizar datos permiten analizar poblaciones pequeñas o solo individuos, pues estás técnicas muchas veces dependen del observador, en cuyo caso, un psicólogo es normalmente el encargado de analizar una muestra de datos de un individuo y determinar con estudios y en base a la experiencia las características de dicho individuo. En el artículo de Markowetz publicado en 2014 nos hablan de un enfoque basado en los datos generados a partir de la tecnología. Los dispositivos tecnológicos resultan ser una buena fuente de información para determinar el comportamiento de los individuos y descubrir patrones en los mismos que se relacionan según el uso de los dispositivos y los datos recolectados. Por lo tanto, se pone hincapié en que gracias a estos recursos que generan cantidades enormes de datos, se puede determinar de cierta manera comportamientos estudiados por la psicología en grupos más pequeños de individuos; puesto que la computación nos permite con técnicas avanzadas de procesamiento de datos, hallar en cortos períodos de tiempo resultados sorprendentes que a mano tomarían un tiempo de meses o incluso años. En el artículo se hizo un estudio para la hipótesis planteada donde se observa a dos clases de individuos, unos con depresión y otros con adicción al Internet. Para el primer grupo se creo una App en Android 4.0 para recolectar información acerca de sus actividades y así poder determinar la energía que empleaban dichos individuos al realizarlas, y si estas actividades estaban ligadas entre si. Aquí el *big data* entraría en juego ayudando a reconocer patrones de comportamiento de la depresión, registrando los datos y analizándolos rápidamente. En un segundo estudio se analizaron los datos de las personas dependiendo de cuánto tiempo pasaban en una red social, se llegó a la conclusión que relacionar el uso del *big data* con mediciones neurocientíficas debería llevar a descubrir un nuevo campo de investigación [18].

Existe en el mundo un campo relacionado a investigar el conocimiento generado en la recolección de datos de la salud, con el fin de cuidar de la salud de los pacientes. Hablamos de la informática de la salud, cuyo objetivo es ayudar a avanzar a la humanidad en la comprensión

de la medicina y la práctica de la misma. Existen numerosas áreas en el campo de la informática de la salud, pero las técnicas de *big data* se pueden aplicar a todas ellas y los resultados que estás técnicas ofrecen, ayudan a reducir costos en investigación, análisis y diagnósticos de enfermedades. Según el articulo publicado en 2014 de Matthew Herland, tan solo en estados unidos, el uso de la informática de la salud podía ahorrarle al país 450 mil millones de dolares cada año, lo cuál habla mucho de estás técnicas basadas en algoritmos de computación. La informática de la salud se divide en varios campos y en cada uno de estos se aplican diferentes técnicas de *big data*. Por ejemplo, en la bioinformática se aplica el filtro de Bloom para encontrar coincidencias rápidamente. En la informática clínica se utilizan algoritmos de predicción basados en los datos de los pacientes, que le permiten a los médicos tomar decisiones de manera ágil y eficiente. En la informática de la salud pública se utiliza el análisis de datos para obtener una visión médica de los datos recolectados. En el análisis de datos de moléculas, se presentan problemas de altas dimensiones debido a las múltiples características de los datos y por esto se usan algoritmos para espacios no euclidianos en este nivel. En cada rama de la ciencia en donde se recogen datos existen técnicas para procesar estos datos y aplicar análisis masivos de datos. [19].

Para los psicólogos, hacer análisis de *big data* no resulta una tarea fácil, es más, la mayoría de las veces los psicólogos tienen datos suficientes para hacer un análisis manual en una computadora de escritorio. Existen veces en que el *big data* se hace necesario para análisis psicométricos. Por ejemplo, en las pruebas internacionales de educación en donde participan miles de profesores y estudiantes, las pruebas en donde se miden datos específicos a nivel mundial, o páginas web como Twitter en que se recolectan datos de las publicaciones y muchas veces se mira el nivel de aceptación de las publicaciones o la felicidad relativa de las personas de acuerdo a un tema en específico. En un artículo de Mike W. en 2016, explica que los psicólogos, a pesar de que no se ven mucho en el campo del *big data*, estos están en parte preparados para afrontar ciertos retos computacionales, debido a que están preparados en diversas áreas que les permiten no solo entender los datos, sino también realizar análisis estadísticos sobre los mismos. El artículo se centra más que todo en un enfoque básico pero eficiente, que los psicólogos pueden utilizar

para hacer análisis de *big data*, y se propone una metodología llamada Dividir/Analizar/Meta-analizar o SAM por sus siglas en inglés, en donde menciona que los algoritmos de análisis pueden ser fácilmente comprendidos y utilizados por los psicólogos, explica que los datos que requieren análisis de *big data* no caben en memoria de procesamiento, y por esta razón se deben partir los datos en conjuntos de datos más pequeños y a esta parte la llama dividir. Luego sigue analizar estos datos separados, para obtener resultados de todos ellos y esta es la fase de análisis, en donde se usan algoritmos para el procesamiento de los mismos. En la parte de meta-análisis comenta, que una vez el análisis de los datos separados den resultados, estos se combinan con técnicas que permiten, en resumen, comprimir las respuestas que se procesaron en menor tiempo por ser analizadas con técnicas de *big data*. Todo esto, hace que sea claro que la psicología no solo necesite técnicas de análisis de *big data*, sino que los psicólogos también pueden utilizar dichas técnicas [20].

En el artículo escrito por Eric Chen y Sean Wojcik, nos dan una guía muy reciente de herramientas que se pueden utilizar para el análisis de datos. Se describe herramientas como Python, el lenguaje de programación interpretado y su librería Pypi (el repositorio de software con las librerías para Python), también hablan de R, el cual es un entorno de software para análisis estadísticos. Por otro lado, comienzan a describir las fases de como debería ser un análisis *big data* y como trasformar datos no estructurados en datos estructurados que puedan ser más útiles a los psicólogos. Primero, se hace énfasis en que antes de empezar con un proyecto de *big data*, se debe planear como administrar los datos, decidiendo si los datos se van a guardar en el computador o en un sistema especializado de bases de datos. Segundo, la adquisición de datos se puede hacer por diversos medios como el scraping de información de páginas web, o la descarga de bases de datos de diversas fuentes como redes sociales. También se pueden descargar datos disponibles en agencias gubernamentales de estados unidos gracias al proyecto Open Data. Tercero, la parte del pre-procesamiento de datos, o limpieza de los datos, es una parte fundamental, ya que una mala limpieza de los mismos podría acarrear problemas más adelante debido a que, en principio, estos datos adquiridos llegan crudos y no estructurados, esta parte se vuelve fundamental. Por último, está la parte de análisis de datos,

en donde se usan las técnicas para extraer la información valiosa a partir de los datos, en esta parte se describen cuatro técnicas. La primera es la analítica de textos, en que se encuentran diversas técnicas de reconocimiento algorítmico dentro de los textos. La segunda, es la técnica de análisis multimedia donde se analizan, por ejemplo, diferentes patrones dentro de los contenidos multimedia y como estos influyen en el comportamiento humano, utilizando técnicas de reconocimiento de *Machine Learning*. La tercera técnica de análisis es el aprendizaje supervisado, en donde básicamente se conocen los datos y se tiene una variable objetivo, la idea es crear un modelo que se ajuste a dicha variable objetivo dependiendo de las características de los datos. Por ejemplo, dadas las características de perros y caballos, podríamos determinar si el próximo dato que estemos analizando es un perro o un caballo. La cuarta técnica es el aprendizaje no supervisado, y al contrario de la anterior, en esta no se conocen datos previos o datos de entrenamiento, por lo que el objetivo es el de descubrir patrones interesantes dentro de los datos, así se puede sacar información para ser etiquetada [21].

3. MARCO TEORICO Y CONCEPTUAL

3.1. Google Spreadsheets Python API

gspread es una API de Google que se utiliza en Python para administrar hojas de cálculo desde el código, esta librería posee múltiples funcionalidades y se utiliza, en algunos casos, en reemplazo de bases de datos convencionales, por su maniobrabilidad e interacción con múltiples productos de Google. La API utiliza unas credenciales otorgadas por Google APIs, para poder acceder a los documentos o aplicaciones en nuestra cuenta de Google.

El algoritmo 1 muestra un ejemplo del uso de la API en Python.

3.2. Pruebas Estandarizadas

Una prueba o test estandarizado, es una prueba que se aplica porque cumple con ciertos estándares o normas previamente definidas. Estas pruebas siempre se aplican de la misma manera, es decir que quienes hacen la prueba y a quienes se les aplica la prueba deben cumplir con las reglas de la prueba estandarizada. Para que una prueba sea estandarizada, debe cumplir con algunos parámetros. El primero de ellos, es que dicha prueba tenga validez o que mida lo que se quiere medir, pues no se puede hacer por ejemplo, una prueba que mida si una persona es propensa a padecer de cáncer si no se sabe con exactitud, cuáles son los factores que determinan ese padecimiento. El Segundo, es que la prueba sea fiable, es decir que produzca resultados que sean parecidos en diferentes individuos si estas pruebas se aplican bajo las mismas condiciones. Por último, la prueba aplicada debe ser exacta, ya sea respecto a que los instrumentos de medición deban funcionar correctamente o que las preguntas preestablecidas sirvan para que las respuestas estén acordes a la realidad [22].

3.3. Pruebas Psicométricas

Estas pruebas miden de cierta manera la psicología de uno o varios individuos en cuestión, y sirven para reconocer, por ejemplo, a que grupo social pertenece un individuo o que ca-

Script de Python 1 API de Google.

```python
import gspread
from oauth2client.service_account import ServiceAccountCredentials
import pprint

scope = ['https://spreadsheets.google.com/feeds']
# obtenemos las credenciales del archivo json
creds = ServiceAccountCredentials.from_json_keyfile_name('client_secret.json', scope)
# autorizamos las credenciales
client = gspread.authorize(creds)

# abrimos la hoja de calculo
sheet = client.open('16PF APQ (respuestas)').sheet1

# obtenemos todos los datos de la hoja
result = sheet.get_all_records()

# obtenemos los datos de una fila especifica
result = sheet.row_values(3) # complete individual row

# obtenemos los datos de una columna especifica
result = sheet.col_values(2) # one aswer of all people with the question

# obtenemos el dato especifico
result = sheet.cell(2,11).value

# imprimimos los datos de la hoja
#print(result)

# cambiamos un dato especifico
sheet.update_cell(2,11,'Falso')

#imprimimos el valor cambiado
result = sheet.cell(2,11).value
print(result)

# Insertamos una fila a la hoja
row = ["estoy", "actualizando", "una", "hoja", "de", "calculo", "desde", "Python"]
index_row = 4
sheet.insert_row(row, index_row)

# Eliminamos una fila de la hoja
index_row = 4
sheet.delete_row(index_row)

# Obtenemos el numero actual de filas en la hoja de calculo
print(sheet.row_count)
```

racterísticas psicológicas presenta, de acuerdo a eso se realiza un diagnóstico, e inclusive se pueden hacer predicciones, porque estas pruebas miden de acuerdo a las condiciones psicológicas y sociales del entorno del individuo y las características del diagnóstico, en que condiciones psicosociales va a estar más propenso a inmiscuirse [22].

3.4. Pruebas Psicométricas Estandarizadas

En la actualidad existen decenas de pruebas estandarizadas para casi todo tipo de situaciones. En este caso vamos a hablar acerca de 4 pruebas bastante populares, las cuales son, Explora (3.4.1), Bat7 (3.4.2), 16PF APQ (3.4.3) e IPP-R (3.4.4), dichas pruebas se evalúan muchas veces de manera manual. Pero, ¿Qué pasaría si dichas pruebas generasen tantos datos que la cantidad de bytes superara lo que la memoria principal es capaz de soportar en procesamiento?, o ¿qué pasaría si los datos fuesen sencillamente tan grandes, que tener respuesta de ellos con técnicas tradicionales sería poco optimo e ineficiente?, pues se deben, entonces, usar técnicas para el tratamiento de grandes volúmenes de datos como el algoritmo THE CURE (3.11). Por otro lado, la evaluación estadística de estos datos, en últimas, usarían las mismas ecuaciones para el análisis de los datos, por ejemplo, calcular según unas respuestas dadas, cuál es la nota de una materia, o según unas características si una persona es propensa a sufrir de depresión. La diferencia no radicaría en el análisis de los datos, sino en el tratamiento de los mismos, es decir, en la manera de administrarlos.

3.4.1. La Prueba Explora

Esta prueba es un cuestionario para la orientación profesional y vocacional. Esta prueba se puede aplicar en uno o varios individuos de 12 años en adelante, cuya finalidad es orientar a quien realice la prueba, siendo ubicado en alguna de las siguientes áreas académicas:

- Técnico-manual.
- Científico-investigador.

- Artístico-creativo.
- Social-asistencial.
- Empresarial-persuasivo.
- Oficina-administración.

El análisis de esta prueba se hace evaluando cada una de las características y el nivel de interés en ellas relacionadas con los campos profesionales definidos. De acuerdo con las puntuaciones obtenidas en cada uno de los campos profesionales, el individuo será más propenso a pertenecer a un campo profesional o a otro [23].

3.4.2. Las Pruebas BAT7

Las pruebas BAT7 son pruebas de aptitud e inteligencia. En esta prueba se miden en 8 aptitudes cognitivas y 3 inteligencias. Las 8 aptitudes cognitivas son:

- Razonamiento Verbal (V).
- Razonamiento Espacial (E).
- Atención (A).
- Concentración (Con).
- Razonamiento Abstracto (R).
- Razonamiento Numérico (N).
- Razonamiento Mecánico (M).
- Ortografía (O).

Las 3 inteligencias que se miden son:

- Factor General(g).
- Inteligencia Fluida (Gf).
- Inteligencia Cristalizada (Gc).

El objetivo de esta prueba es en sí, la evaluación de las aptitudes cognitivas. Una aptitud cognitiva es la capacidad para adquirir nuevos conocimientos, es decir la capacidad de un individuo de obtener información y procesarla [24].

3.4.3. La Prueba 16PF APQ

Es un test de personalidad que está hecho para evaluar individuos entre los 12 a los 19 años, en esta prueba se miden 16 rasgos de personalidad, y 5 dimensiones globales. Los rasgos de personalidad son llamadas escalas, y estas escalas se miden con 15 cuestionarios que determinan los niveles en los rasgos de la personalidad y determinan el comportamiento del individuo. Los 16 rasgos de personalidad son, afabilidad, razonamiento, estabilidad emocional, dominancia, animación, atención normas, atrevimiento, sensibilidad, vigilancia, abstracción, privacidad, aprensión, apertura al cambio, autosuficiencia, perfeccionismo y tensión; mientras que las 5 dimensiones globales son, extraversión, ansiedad, dureza, independencia y autocontrol [25].

3.4.4. La Prueba IPP-R

La prueba IPP-R es una nueva versión de la prueba IPP por lo que sus siglas significan Intereses y Preferencias Profesionales-Revisado, la finalidad de esta prueba no es la de medir la inteligencia sino de orientar a la persona a escoger una profesión de acuerdo a sus intereses y preferencias las preguntas que se hacen aquí son de dos tipos, las que requieren que el individuo elija entre diferentes profesiones para saber cuál es de su gusto y las que requieren que el mismo escoja entre diferentes tareas o actividades, esta prueba se reparte en un cuestionario de 180 preguntas de ambos tipos de respuesta [26].

3.5. Espacios Euclidianos y Espacios No Euclidianos

Se define como espacio euclidiano como aquel espacio en un plano que sencillamente cumple con las reglas de espacios euclidianos, y los espacios no euclidianos como aquellos que no cumplen o difieren en al menos una regla. Las reglas de los espacios euclidianos son:

1. "Dos puntos cualesquiera determinan un segmento de recta".
2. Ün segmento de recta se puede extender indefinidamente en una linea recta".
3. "Se puede trazar un centro y una circunferencia dados un centro y un radio cualquiera"
4. "Todos los ángulos rectos son iguales entre si".
5. "Si una linea recta corta a otras dos, de tal manera que la suma de los dos ángulos interiores del mismo lado sea menor que dos rectos, las otras dos rectas se cortan, al prolongarlas, por el lado en el que están los ángulos menores que dos rectos". "Por un punto exterior a una recta, se puede trazar una única paralela".

En cambio, los espacios no euclidanos pueden ser curvos, como elipses o hipérbolas, de esta forma los métodos de medición de distancias para los espacios euclidianos y los espacios no euclidianos difieren [27].

3.6. Clustering

Es un método mediante el cuál podemos separar en grupos llamados clusters distintos tipos de datos, dichas agrupaciones sirven para identificar, precisamente, datos con propiedades especificas que difieren entre clusters.

Usando métodos para el descubrimiento de clusters, podemos optimizar el tiempo de búsquedas y actualización de los mismos. Se vuelve necesario utilizar estos métodos en bases de datos masivas en donde la cantidad de datos requiere un buen desempeño en el procesamiento de los mismos.

Existen diferentes tipos de mediciones de distancias para espacios euclidianos, aunque también existen mediciones de distancia para espacios no euclidianos o espacios euclidianos de grandes dimensiones. Podemos usar técnicas de agrupamiento tanto en espacios euclidianos como en espacios no euclidianos.

Para armar bien nuestros clusters debemos hacer una serie de pasos en los cuales lo más importante es medir las distancias. Existen varias distancias que podemos medir, las cuales son la distancia de cada punto a su centroide o clustroide, la distancia intercluster y la distancia intracluster.

Para determinar los clusters, debemos utilizar distancias diferentes para cada tipo de espacio, por ende, las distancias para espacios euclidianos no son las mismas que para los espacios no euclidianos, por un lado, para los espacios euclidianos tenemos algunas distancias como:

- La distancia euclidiana común o distancia ordinaria.
- La distancia Manhattan o longitud Manhattan.
- La distancia de Chebyshov.

Por otro lado, hay algunas distancias para espacios no euclidianos como [28]:

- La distancia Jaccard o índice de Jaccard.
- La distancia Cosine o similitud coseno.
- La distancia de Hamming.
- La distancia de edición o distancia Levenshtein.

3.6.1. El Centroide

El centroide es un punto que se encuentra en el centro de un cluster de datos, normalmente definido por el promedio de la suma de las coordenadas de un conjunto de datos, usamos el centroide cuando nos encontramos en espacios euclidianos, el centroide puede ser un punto imaginario [29].

3.6.2. El Clustroide

El clostroide se determina seleccionando de un conjunto la suma mínima de sumas de los cuadrados de las distancias de los puntos en el cluster.

Usamos el clostroide cuando estamos en espacios no euclidianos, ya que allí no hay concepto de punto medio, puesto que el promedio de la suma de las coordenadas en el cluster debe ser un punto no imaginario presente en los datos [29].

3.6.3. La Distancia Inter-cluster

La distancia intercluster es la distancia que hay entre los diferentes cluster, cuando aplicamos técnicas de Clustering, esta distancia tiende maximizarse [30].

3.6.4. La Distancia Intra-cluster

La distancia intracluster es la distancia que hay entro los diferentes datos dentro de un mismo cluster en cada uno de los clusters, al contrario de la distancia intercluster, esta distancia tiene a minimizarse, es decir, que los datos dentro de un cluster se juntan cada vez más en vez de separarse [30].

3.6.5. Distancia Euclidiana Común

Esta distancia se calcula con respecto a al centroide previamente definido en cada cluster. Se calcula la distancia euclidiana para saber a que cluster pertenece un dato especifico, y se hace lo mismo con todos los datos. El dato se queda en el cluster al que su distancia respecto al centroide del cluster sea menor. La distancia euclidiana, se calcula después de calcular el centroide de los clusters, por lo que la distancia euclidiana entre las coordenadas de un dato $D = (d1, d2)$ y el centroide de un cluster $C = (c1, c2)$, sería $d_E(D, C) = \sqrt{(d1 - c1)^2 + (d2 - c2)^2}$ [31].

3.6.6. Distancia Manhattan

La distancia Manhattan es una distancia más realista si se aplica a la vida real, esta se llama así porque casi todas las cuadras de la ciudad de Manhattan son en forma de cuadricula, y puesto que es mas realista medir las cuadras por la longitud de los cuadros de la cuadricula que en linea recta atravesando los edificios, entonces cualquier camino que se tome de un punto a otro es exactamente el mismo, pues el camino en linea recta es imposible en esta distancia, por lo que la formula que representa la medición de esta distancia, al igual que la medición de la distancia euclidiana, toma en cuenta dos puntos, en donde un punto $D = (d1, d2)$ son las coordenadas de un dato y otro punto $C = (c1, c2)$ son las coordenadas de un centroide, la distancia se determina entonces por, $d_M(D, C) = |d1 - c1| + |d2 - c2|$, es decir, la suma de las partes enteras de las diferencias [31].

3.6.7. Distancia de Chebyshov

La distancia Chebyshov también es llamada la distancia del tablero de ajedrez, y esto se debe a que la distancia se toma midiendo como se movería la pieza del rey en el tablero de ajedrez, es decir, un solo paso en todas direcciones, es decir que a diferencia de la distancia Manhattan, esta no se restringe solamente a medir en horizontal y en vertical, sino también en diagonal, la formula en donde el punto $D = (d1, d2)$ y $C = (c1, c2)$, sería la siguiente, $d_C(D, C) = max(|c1 - d1|, |c2 - d2|)$ [31].

3.6.8. Distancia Jaccard o Índice de Jaccard

La distancia Jaccard o Indice de Jaccard, es una distancia que mide el nivel de similitud entre dos conjuntos, este nivel de similitud es un valor entre 0 y 1, esto indica que cuando más se acerque el indice a 1 mas similar es, siendo 1 el valor máximo el cuál indica que ambos conjuntos son idénticos, y 0 el valor mínimo que indica que ningún item del conjunto A tiene una instancia en el conjunto B, la fórmula sería la siguiente donde $A = (n1, n2, n3, n4)$ y

$B = (n1, n2, n3, n4)$ la distancia Jaccard se calcula así $d_J = |A \cap B|/|A \cup B| = 0$, que quiere decir, la división de las cardinalidades de la intersección y la unión de ambos conjuntos [31].

3.6.9. Distancia Coseno o Similitud Coseno

La distancia Coseno o similitud coseno, se mide para saber el nivel de similitud entre dos vectores gracias al ángulo existente entre ellos, es decir que si el valor resultante es 0 exacto, no existe un ángulo entre ambos vectores por lo tanto los vectores son idénticos, en caso contrario, si el valor es diferente de 0 y oscila entre 1 y -1, los vectores no son iguales, la fórmula donde $A = (n1, n2, n3, n4)$ y $B = (n1, n2, n3, n4)$ sería la siguiente,

$d_C = (\sum_{i=1}^{N} A_i * B_i)/(\sqrt{\sum_{i=1}^{N} A_i^2} * \sqrt{\sum_{i=1}^{N} B_i^2})$ [31].

3.6.10. Distancia Hamming

Esta distancia solo se puede usar entre cadenas de igual longitud, consiste en determinar cuanto se diferencia una cadena o vector de otro, por ejemplo: Considerando $A = (1, 0, 1, 0, 1, 0)$ y $B = (0, 0, 1, 1, 1, 0)$ la distancia Hamming entre ambos vectores es 2, pues en las posiciones 0 y 3 sus valores difieren entre si, $d_H(A, B) = 2$, esta distancia es igual a 0 solo si ambos vectores son idénticos [31].

3.6.11. Distancia edición o Distancia Levenshtein

Esta distancia también se conoce como distancia entre palabras, a diferencia de la distancia Hamming, en esta distancia se pueden usar cadenas de diferente longitud, consiste en determinar cuantas modificaciones se le debe hacer a una cadena o vector para ser transformada en otra,usando el numero mínimo de modificaciones necesarias por ejemplo, siendo $A =$ "*Avion*" y $B =$ "*Accion*", nótese primero que la longitud de ambas cadenas son diferentes y el número de modificaciones sobre la cadena A para ser transformada en B seria de 2, pues en la cadena B se cambia la v por una c y se añade otra c [31].

3.7. Data Science

Data Science o ciencia de datos es un término que engloba un conjunto de herramientas, técnicas y soluciones. El fin último es el de generar conocimiento valioso a partir de grandes cantidades de datos. La forma de llegar a ese conocimiento es a partir de la interpretación de los datos y la correcta comunicación de los mismos. La ciencia de datos utiliza cosas como el Machine Learning, las matemáticas, la estadística y altos conocimientos informáticos, para la extracción, procesamiento y visualización de resultados de cualquier conjunto de datos. La información obtenida puede representar un alto valor para el mercado o la sociedad. Por lo tanto es un término que engloba distintas disciplinas y requiere además del correcto entendimiento de la población analizada, ya sea un nicho de mercado, una problemática social o un propósito científico [32].

3.8. Big Data

Big Data o Datos masivos es un concepto que hace referencia a grandes cantidades de información que no puede ser procesada con métodos de análisis convencionales debido a su gran tamaño. Este tipo de información puede ser recopilada de distintas fuentes y en distintos formatos, pero muchas veces la cantidad de los mismos impiden a las máquinas de cómputo que los datos puedan ser analizados. Esto es debido a la capacidad limitada de procesamiento que poseen. Por esta razón, se implementan técnicas que reducen el tiempo de procesamiento y permiten a su vez el tratamiento de enormes bases de datos. La ventaja del *big data*, es que los resultados que nos entrega el analizar datos masivos, nos permite una mayor aproximación a nivel estadístico que si se analizaran los datos en menor medida. Por ende, dichos resultados se acercan más a la realidad, de esta manera se proporcionan resultados más confiables respecto a miles de cuestiones. El objetivo del *big data* es convertir los datos en información para la toma de decisiones en tiempo real, corto o largo plazo. El análisis con las técnicas de procesamiento de *big data* es tan importante, que a menudo a los datos analizados se les ha denominado .[el] petróleo del futuro". Existen cantidad de ejemplos del uso del *big data* en la vida real, en la

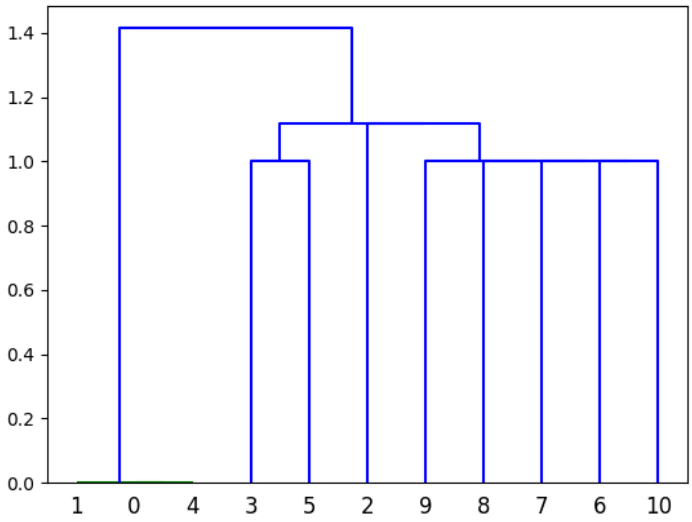

Figura 1. Dendrograma

salud para la prevención de virus como el H1N1, en la seguridad para la prevención de delitos, en la política para la toma de decisiones e incluso para conocer que palabras utilizar en los discursos políticos, en las aerolíneas para saber cuando comprar pasajes de avión, etc [32].

3.9. Clustering Jerárquico

El clustering jerárquico es un método de agrupamiento de datos en forma jerárquica, en donde se agrupan datos con características semejantes. Se empieza con la idea de que todos los datos son sus propios centroides o clustroides (esto dependiendo si el espacio es euclidiano o no), estos datos se agrupan posteriormente con otros datos cuya distancia sea la mas mínima formando un cluster de dos o mas datos. De esta manera, los clusters van creciendo dependiendo de la distancia de los datos hasta alcanzar la cantidad de clusters deseados. Un solo cluster representa una mayor cantidad de procesamiento, mientras que el menor procesamiento significa tener n clusters igual a la cantidad de los datos que se tienen [33].

El algoritmo 2 es un ejemplo de como se implementa clustering jerárquico en Python utilizando el paquete scipy. La figura 1 muestra el dendrograma del agrupamiento resultando.

Script de Python 2 Clustering Jerárquico.

```
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
import numpy as np

# Datos
a = np.array([[1,   0  ],
  [1,   0  ],
  [1,   1.5  ],
  [0,   1  ],
  [1,   0  ],
  [0,   2],
  [2,   1],
  [2,   2  ],
  [2,   3  ],
  [3,   2  ],
  [3,   3  ]])

# Se hace el clustering jerarquico
z = linkage(a)

# Se crea el dendrograma
d = dendrogram(z)

# Se imprime el dendrograma
plt.show()
```

El resultado muestra como se agrupan primero los datos cuyas distancias sean mínimas y posteriormente se van juntando clusters dependiendo de las distancias entre ellos.

3.10. K-Means

El algoritmo K-Means es un algoritmo de agrupamiento de datos con características semejantes, este método de agrupamiento requiere como parámetros el número K de clusters deseados y un número K de centroides dados, si los centroides no se dan se escogen aleatoriamente. Se empieza con K centroides distribuidos en los datos y se halla la distancia de todos los demás puntos a los centroides, con el objetivo de determinar a que cluster pertenecen dichos puntos, esto quiere decir que los puntos formarán parte del cluster del centroide al que hayan tenido la menor distancia, una vez se hayan formado los K clusters, se calcula entre los puntos de cada cluster un nuevo centroide. Un centroide significa un punto que se encuentre en medio de los datos del cluster, posteriormente se vuelve a calcular los clusters, pues algunos puntos pueden o no quedar más cerca de los nuevos centroides, esto indica que algunos puntos cambian de cluster, todo el procedimiento se repite hasta la convergencia, es decir, hasta que no haya más intercambio de datos entre clusters (grupos de datos) [34].

El algoritmo 3 es un ejemplo de como se implementa KMeans en Python Este es ejemplo de cómo se implementa k-means en Python. La figura 2 muestra como se vería el resultado

Como se puede observar al pasarle como parámetro a KMeans el número de clusters, el código genera 3 clusters, representados gráficamente en 3 colores diferentes, con sus respectivos centroides.

Los centroides no representan puntos de datos reales en el algoritmo, para espacios no euclideanos se necesitan clostroides que son puntos reales en el espacio de datos, debido a las múltiples características de los datos.

Script de Python 3 K Means.

```
from matplotlib import pyplot as plt
from sklearn.cluster import KMeans
import numpy as np

# Datos
a = np.array([[1,   0  ],
  [1,   0  ],
  [1,   1.5  ],
  [0,   1  ],
  [1,   0  ],
  [0,   2],
  [2,   1],
  [2,   2  ],
  [2,   3  ],
  [3,   2  ],
  [3,   3  ]])

# se le envian los parametros para calcular kmeans
kmeans = KMeans(n_clusters=3)
# La funcion fit calcula kmeans con los datos
kmeans.fit(a)

# Se obtienen los labels asignados
labels = kmeans.labels_
# Se obtienen los centroides
centroids = kmeans.cluster_centers_
# posible rango de colores por cluster
colors = ["g.","r.","c.","b.","k.","o."]

# recorre cada dato y lo imprime a color segun al cluster en que
    pertenezca
for i in range(len(a)):
  plt.plot(a[i][0], a[i][1], colors[labels[i]])

# dibuja la ubicacion de los centroides
plt.scatter(centroids[:,0],centroids[:,1], marker='x')
# muestra el grafico generado
plt.show()
```

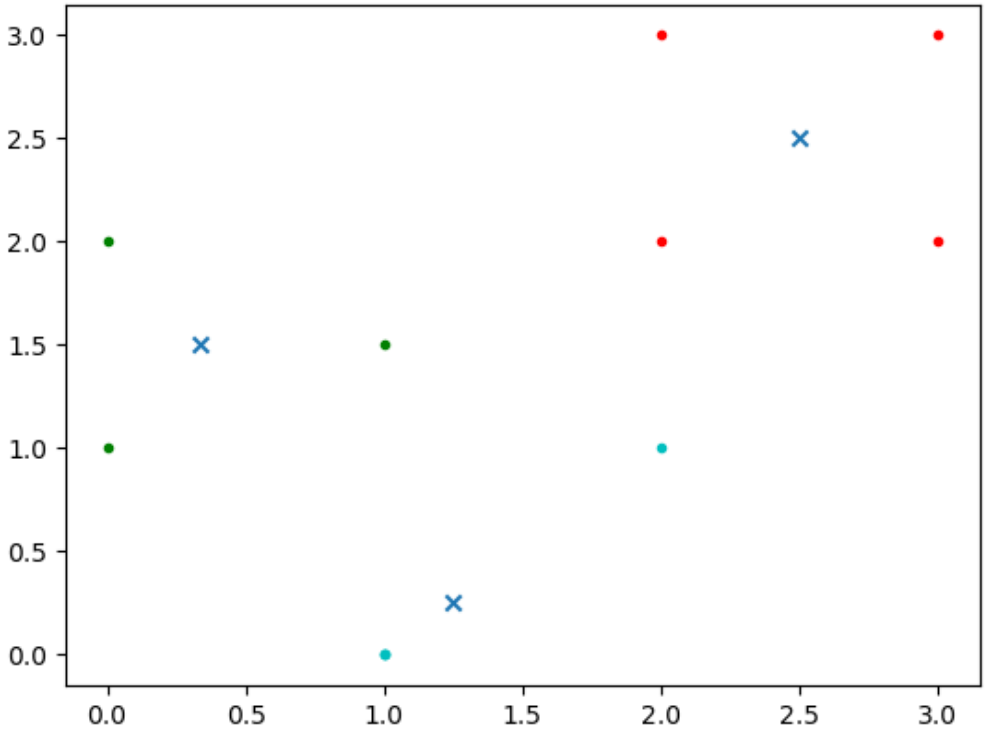

Figura 2. Clusters con k-means

3.11. THE CURE

El algoritmo THE CURE es una solución cuando se trata de hablar del análisis de datos de manera masiva, datos que no caben en la memoria ram y por ende su procesamiento es imposible, estos datos tan grandes requieren de técnicas como las que aplica THE CURE, es decir el de escoger unos datos o puntos que representen a los clusters, para que puedan ser procesados en un tiempo considerablemente bajo para el tamaño de los datos que se analizan. Este algoritmo tiene varias fases. Primero, se tienen todos los datos y se toma una muestra de los datos lo suficientemente pequeña para caber en la memoria de procesamiento. Segundo, se aplica el algoritmo de clustering jerárquico hasta obtener K cantidad de clusters previamente predeterminados. Tercero, se escogen los puntos más representativos de cada uno de los clusters. Una forma de entender que son estos estos puntos, es que son como si tuviéramos varios centroides dentro del cluster y estos se escogen lo más alejados entre si como sea posible. Cuarto, se mueven los puntos representativos hacia el centroide del cluster en un porcentaje alfa que se haya definido Quinto, se actualizan los clusters, tomando cada punto p y agrupándolos al cluster del punto representativo más cercano a estos. Por último se repite el proceso hasta que convergan todos los puntos en cada cluster, recalculando en cada paso los puntos representativos, los centroides y los datos en los cluster [34].

3.12. Guia de Uso del Usuario

En esta sección se explica como un usuario común puede hacer uso de esta herramienta, lo que le permitirá acelerar los procesos de psicoanálisis de poblaciones enormes, separando los grupos de individuos con características similares de manera automática, para este objetivo lo primero que necesitará el usuario es un formulario de preguntas psicométricas hecho en Google Forms y que no contenga respuestas abiertas, con el fin de que los individuos respondan. De esta manera, se genera automáticamente una hoja de respuestas en Google Sheets, el formulario de preguntas se crea de la siguiente manera.

1. El usuario debe tener una hoja de cálculo con el test de análisis psicométrico organizando las filas de forma ordenada, donde la primera columna es la pregunta y las demás columnas de la fila son las respuestas. El archivo debe estar guardado en formato .csv, este archivo debe ir dentro de la carpeta adjunta con el código de la tesis.

	A	B	C	D
1	1. Aunque haya tenido ideas mejores que mis amigos, me quedo tranquilo y no las digo.	Verdadero	?	Falso
2	2. Me gusta sorprender a los demás haciendo alguna gracia o diciendo cosas divertidas para ellos.	Verdadero	?	Falso
3	3. Me cuesta bastante hablar con alguien que no conozco.	Verdadero	?	Falso
4	4. Me gustaría más llegar a ser artista o escritor que oficial de la policía.	Verdadero	?	Falso
5	5. Yo diría que se puede confiar en la mayoría de las personas que conozco.	Verdadero	?	Falso
6	6. Normalmente me considero una persona realista, que atiende más a los aspectos prácticos.	Verdadero	?	Falso
7	7. Hablo de mis sentimientos si alguien se interesa por ellos.	Verdadero	?	Falso
8	8. A veces me siento culpable o deprimido, aun cuando no haya hecho nada malo.	Verdadero	?	Falso
9	9. Me gusta la comida o los platos ya conocidos, más que los nuevos o especiales.	Verdadero	?	Falso
10	10. Si tuviera unos minutos libres, me gustaría emplearlos estando yo solo y tranquilo.	Verdadero	?	Falso
11	11. A veces, aunque haya sido durante un tiempo breve, he tenido sentimientos de odio hacia mis padres.	Verdadero	?	Falso
12	12. Normalmente estoy relajado, aun cuando me toque esperar por algo.	Verdadero	?	Falso
13	13. Cuando algo me abruma o me molesta, normalmente me recupero con facilidad.	Verdadero	?	Falso
14	14. Cuando alguien me interrumpe mientras estoy diciendo algo, cedo y le dejo de hablar.	Verdadero	?	Falso
15	15. Se pueden romper un poco las normas si ninguna persona recibe daño por ello.	Verdadero	?	Falso
16	16. Cuando e incorporo a un grupo nuevo, me mantengo en un segundo plano durante un tiempo.	Verdadero	?	Falso

Figura 3. Archivo del formulario con opciones de respuesta

2. El usuario se debe dirigir a google drive desde la siguiente dirección https://drive.google.com y crea una nueva carpeta dando clic a nuevo, carpeta y luego en crear.

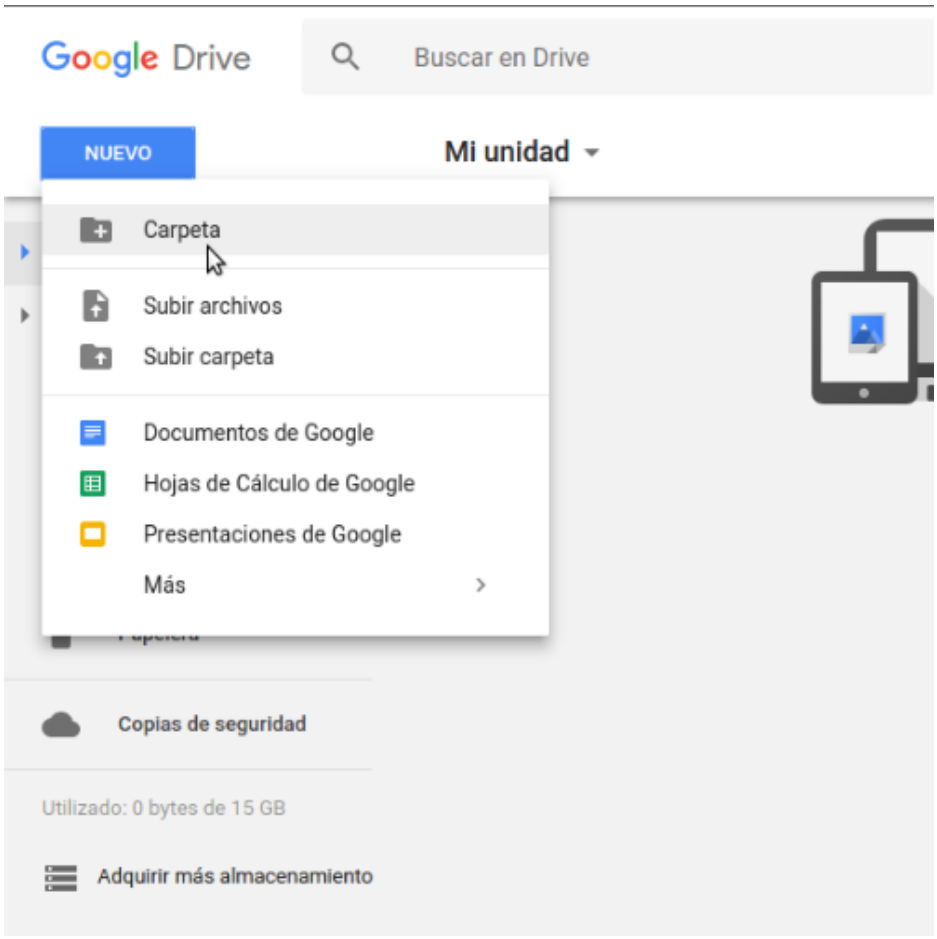

Figura 4. Creación de una carpeta en Google Drive

3. El usuario debe entrar a la nueva carpeta recién creada y da clic en nuevo, más, formularios de google.

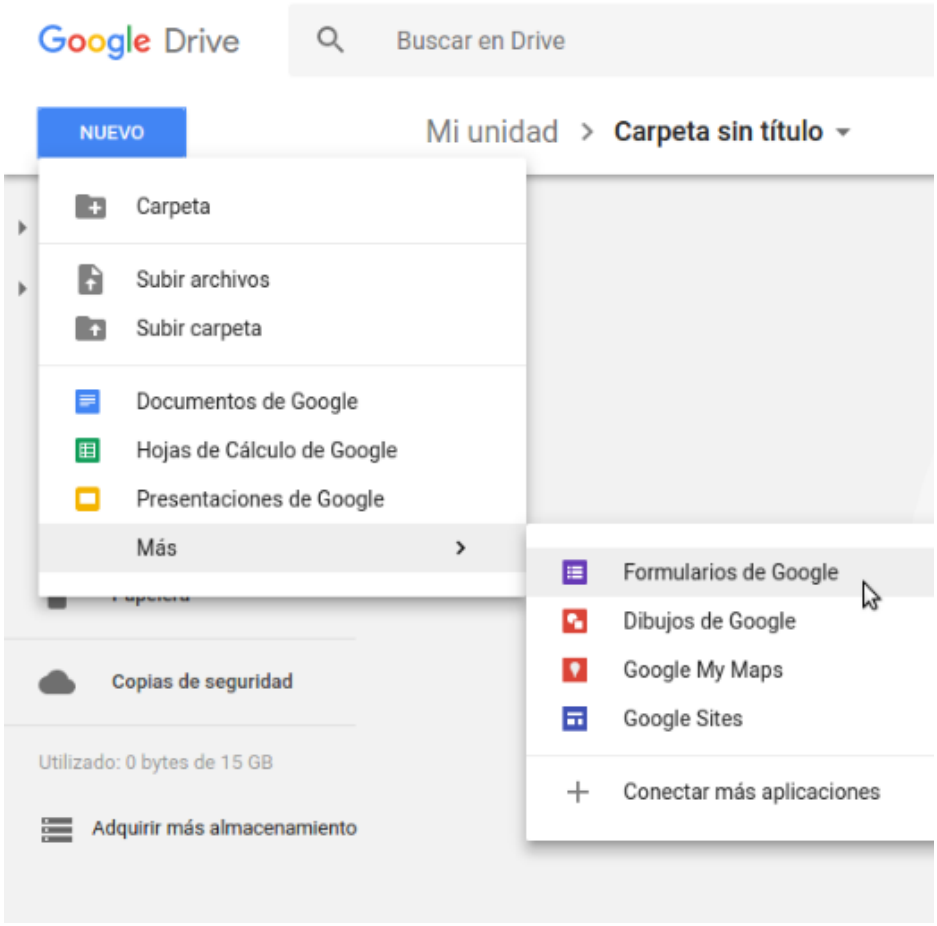

Figura 5. Creación de un Formulario de Google

4. El usuario debe escribir donde dice Formulario sin título, el nombre de la prueba psicométrica, y luego crea en orden lo siguiente: primero, la opción para que el individuo escriba el nombre. Segundo, la opción para que escriba el documento de identidad. Tercero, las preguntas y sus respectivas opciones de respuesta. El usuario puede añadir cosas como imágenes, videos, descripción de la prueba o de las preguntas y nuevas secciones, pero las preguntas deben ser de opción múltiple.

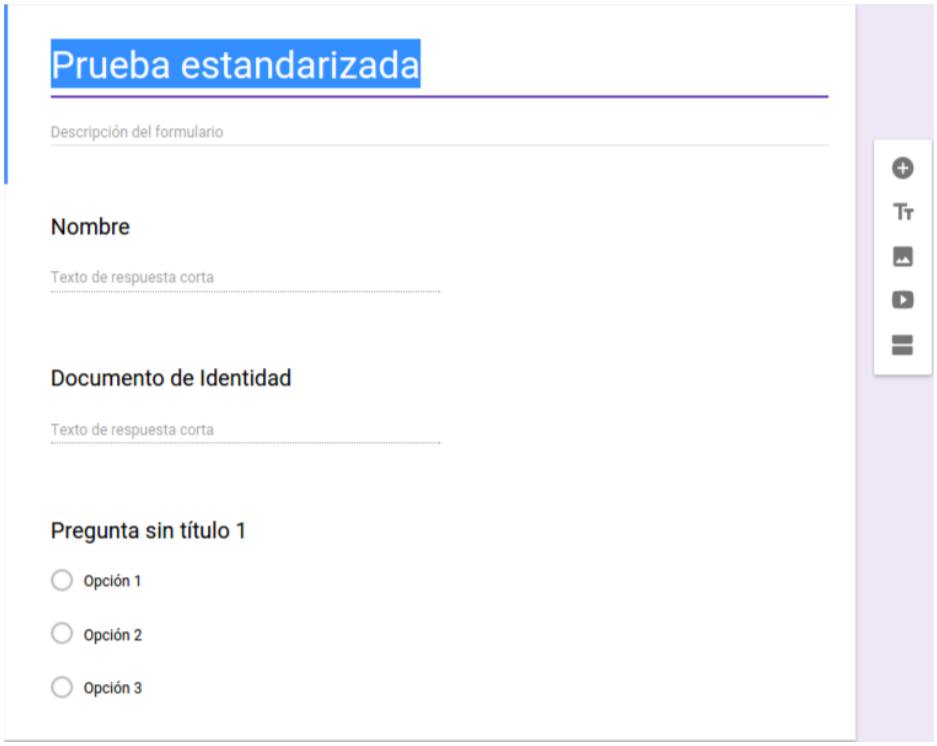

Figura 6. Creación de una prueba online

5. Una vez finalizado de escribir todas las preguntas con sus opciones de respuesta, el usuario debe tener en el computador el sistema operativo linux, y la herramienta pip instalada. Para instalar pip, debe presionar a la vez ctrl+alt+T, a continuación se muestra una consola de comandos en la pantalla, donde el usuario ingresa el comando sudo apt-get install Python-pip y presiona la tecla enter.

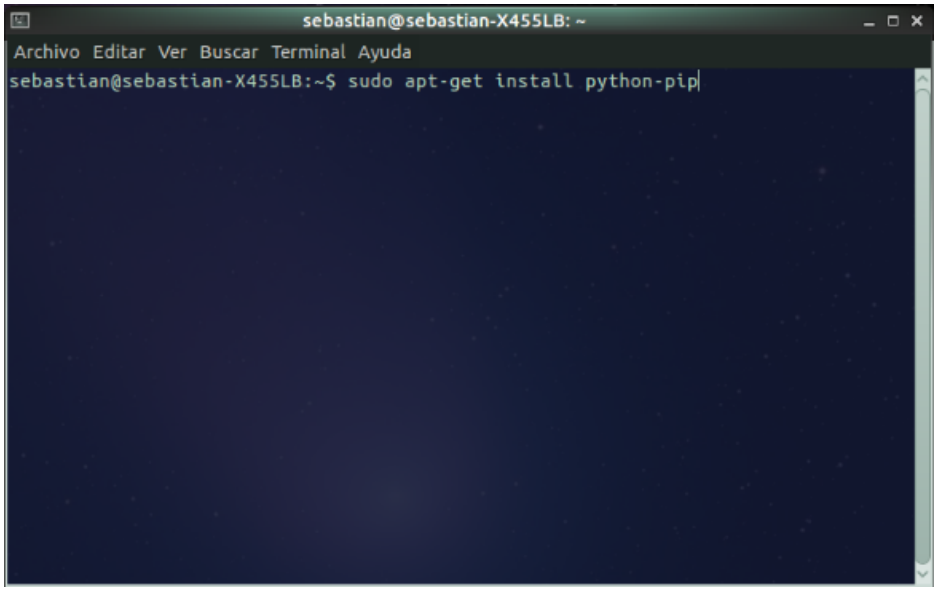

Figura 7. Instalación de pip

6. El usuario debe tener en su computador la carpeta adjunta con el código de la tesis del proyecto de grado, en esta misma carpeta debe estar ubicado el archivo del paso 1, a continuación se debe dirigir a
la carpeta donde se encuentra el código, dar clic derecho y luego dar clic en abrir en un terminal

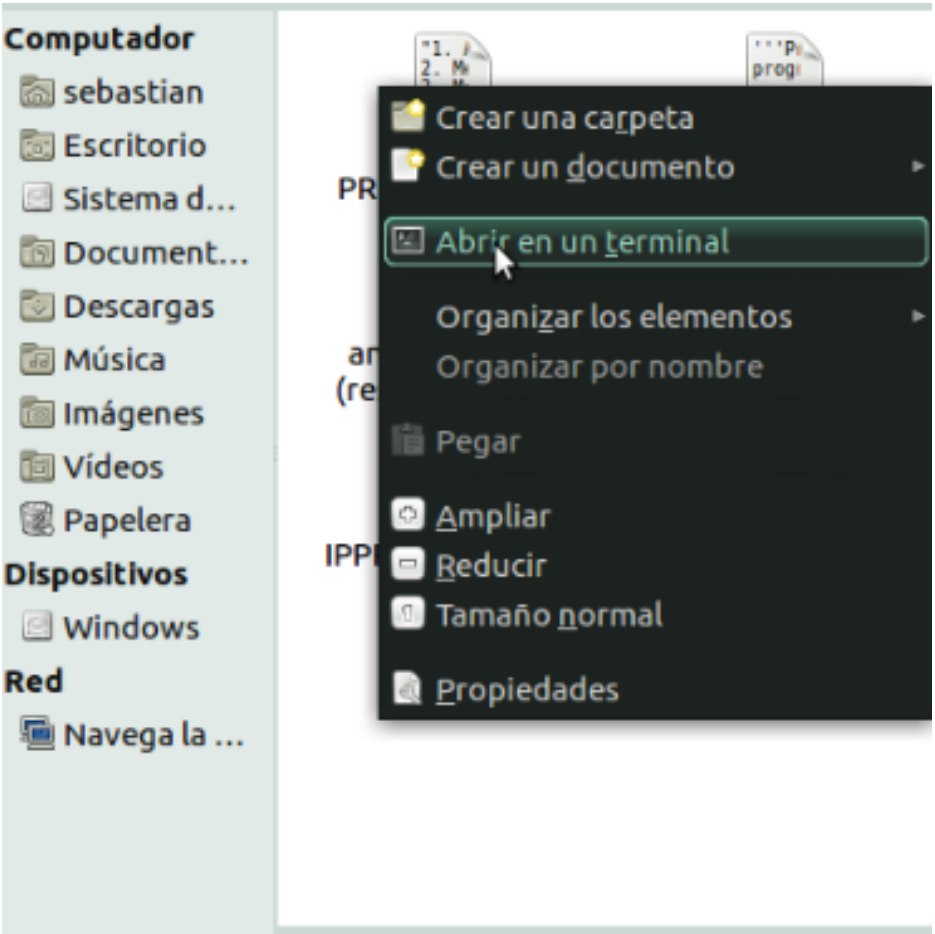

Figura 8. Abrir una terminal con el archivo

7. Se abrirá una terminal ubicado en la carpeta actual del proyecto, el usuario debe escribir el comando sudo pip install -r requirements.txt y dar
enter, esto instala los paquetes de Python necesarios para que el código funcione correctamente.

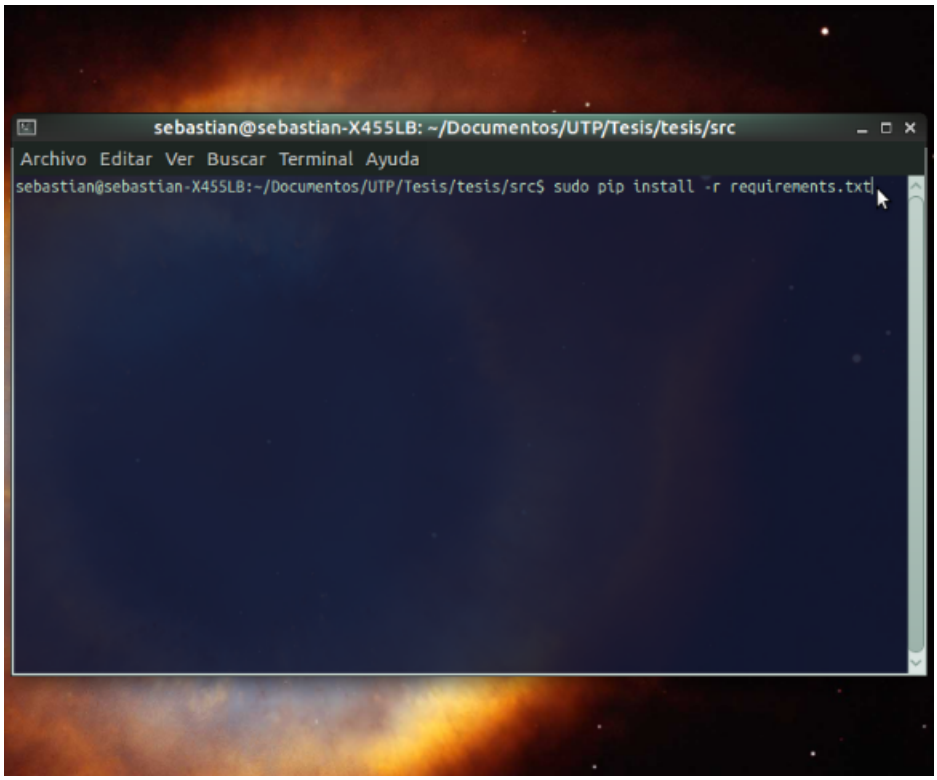

Figura 9. Instalación de Requisitos

8. A continuación el usuario debe abrir el archivo client_secretjson dando clic derecho sobre el mismo y seleccionando Abrir con, y luego Editor de textos

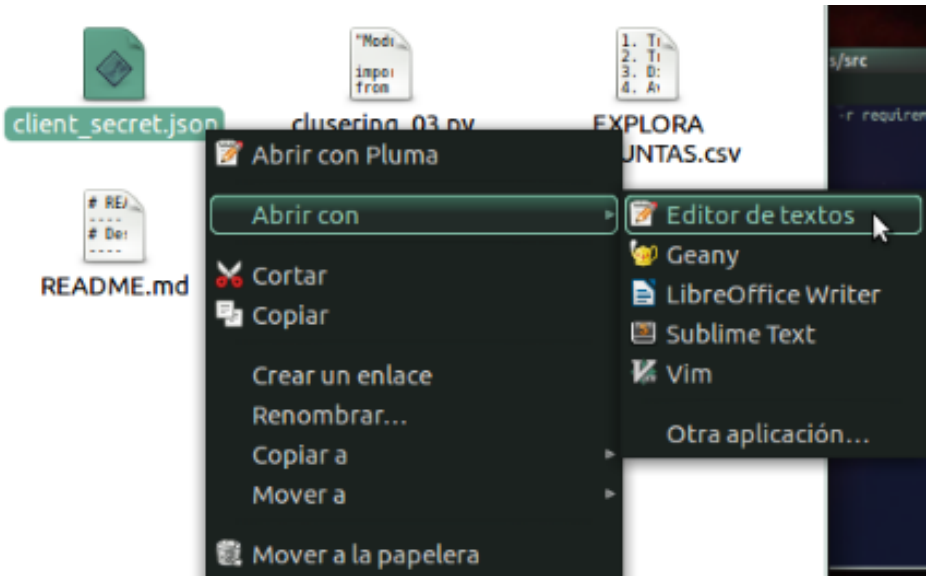

Figura 10. Archivo cliente_secret

9. Se abre un código en un editor de textos donde
se debe seleccionar y copiar el correo que se encuentra donde dice client_email, sin las comillas.

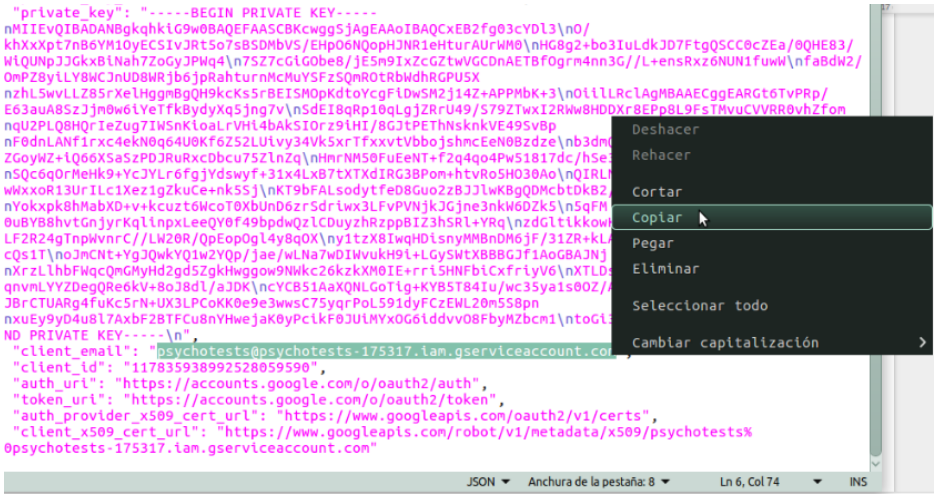

Figura 11. Correo del cliente

10. Una vez copiado el correo se debe entrar dentro de la prueba de análisis creada en el paso de la figura 4, ir a respuestas y dar clic al botón verde (crear hoja de cálculo) y luego en crear.

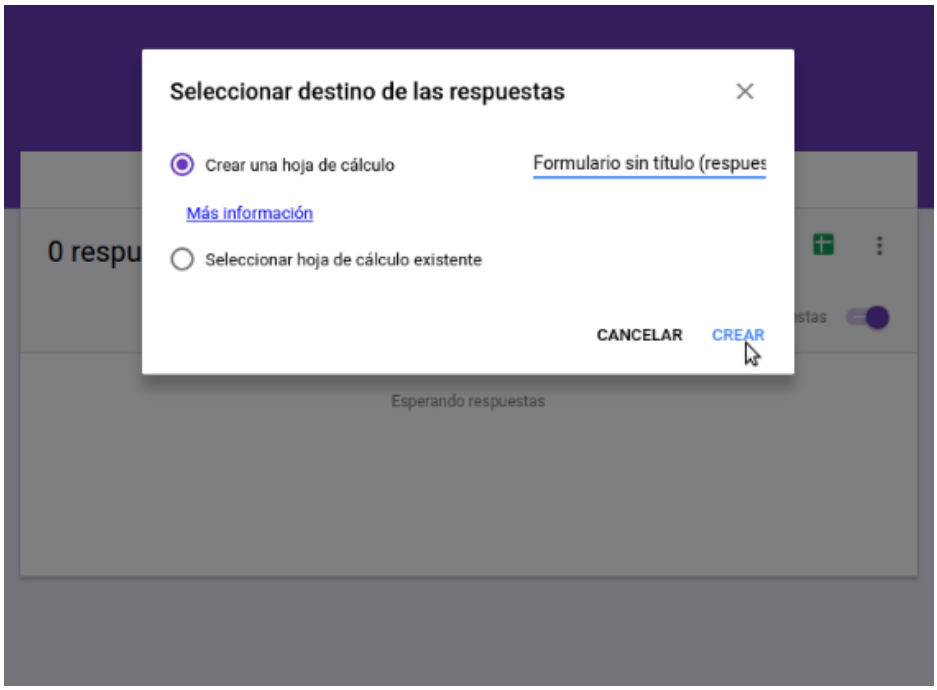

Figura 12. Creación de Hoja de Cálculo

11. Se debe dar clic nuevamente al botón verde, que ahora se llama Ver respuestas en Hojas de cálculo. Se abrirá una nueva pestaña con una hoja de cálculo donde estarán todas las respuestas de quienes respondan el formulario o test de análisis. Una vez dentro de la hoja de cálculo, se le da clic al botón compartir y allí se pega el correo previamente copiado en el paso de la figura 11. Por último, se le da clic en enviar, estos pasos le dan permiso al código del proyecto de extraer todas las respuestas en un formato legible para el código.

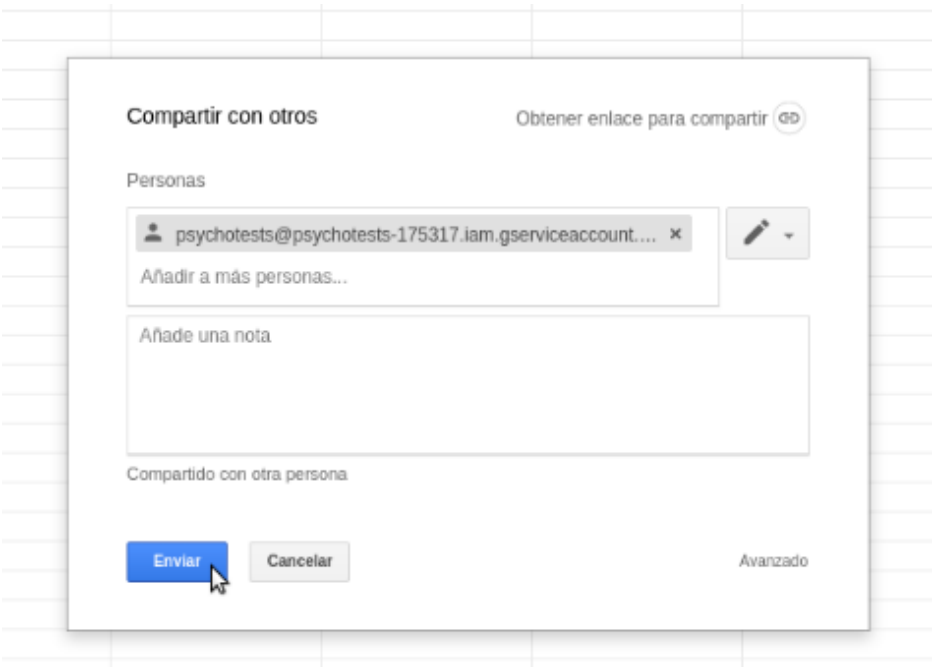

Figura 13. Asignación de Permisos

12. Una vez obtenidas las respuestas de los pacientes, el usuario debe ir a la carpeta del proyecto dentro del computador, dar clic derecho sobre el archivo answerspy y luego en Abrir con, Editor de textos, allí se debe dirigir a la función principal llamada main(), dentro de dicha función se encuentran dos variables que hacen referencia a varios archivos, estas dos variables son answer_options_sheets que hace referencia a tipos de archivos como el creado en el paso de la figura 3, y answer_sheets que hace referencia a tipos de archivos que contienen respuestas de formularios creados en google drive.

```
def main():
    "Main Function"


    answer_options_sheets = [
        '/home/sebastian/Documentos/UTP/Tesis/tesis/src/16PF APQ PREGUNTAS.csv',
        '/home/sebastian/Documentos/UTP/Tesis/tesis/src/EXPLORA PREGUNTAS.csv',
        '/home/sebastian/Documentos/UTP/Tesis/tesis/src/IPPR PREGUNTAS.csv']

    answer_sheets = [
        '16PF APQ (respuestas)',
        'EXPLORA (respuestas)',
        'IPP-R (respuestas)']
```

Figura 14. Variables de la función Principal

13. Se debe entonces añadir en answer_options_sheets, la dirección donde se encuentra ubicado el archivo creado en el paso de la figura 3, para esto, se debe dirigir a la carpeta donde se encuentra dicho archivo, la cual es la misma carpeta del proyecto donde se encuentra el código, se le da clic derecho sobre el archivo y luego en Propiedades, esto abre un recuadro con la información del archivo donde interesan dos cosas, el Nombre y la Ubicación.

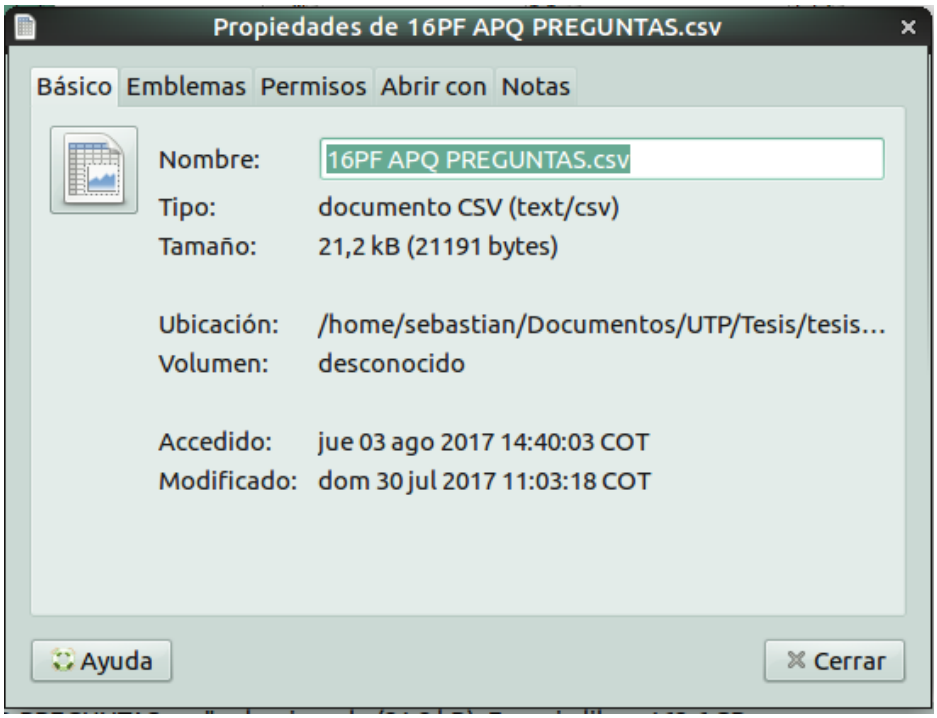

Figura 15. Nombre y Ubicación

14. Se debe copiar la ubicación y luego pegarla dentro de la variable answer_options_sheets vista en el paso de la figura 15 y seguida de un "/", luego se copia el nombre y se pone después del "/", como en este ejemplo, queda ubicado en la primera posición.
se pega la ubicación dentro de comillas simples

```
answer_options_sheets = [
    '/home/sebastian/Documentos/UTP/Tesis/tesis/src/
    '/home/sebastian/Documentos/UTP/Tesis/tesis/src/EXPLORA PREGUNTAS.csv',
    '/home/sebastian/Documentos/UTP/Tesis/tesis/src/IPPR PREGUNTAS.csv']
```

Figura 16. Asignación de la Ubicación

se debe pegar
el nombre después del "/" y después cerrar las comillas simples, por ultimo se termina con coma

```
answer_options_sheets = [
    '/home/sebastian/Documentos/UTP/Tesis/tesis/src/16PF APQ PREGUNTAS.csv',
    '/home/sebastian/Documentos/UTP/Tesis/tesis/src/EXPLORA PREGUNTAS.csv',
    '/home/sebastian/Documentos/UTP/Tesis/tesis/src/IPPR PREGUNTAS.csv']
```

Figura 17. Asignación del nombre de las preguntas

15. Ahora debe di-
rigirse a la hoja de respuestas creada en el paso de la figura 12 y copiar el nombre de dicha hoja

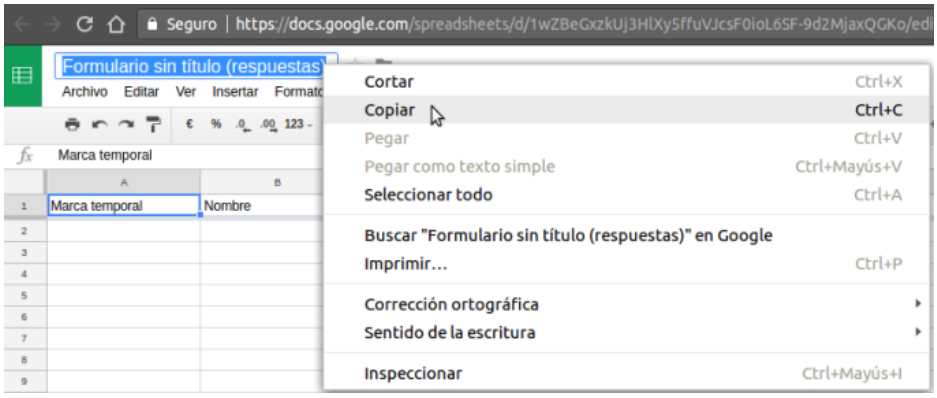

Figura 18. Selección del nombre de la hoja

16. Se debe pegar dicho nombre dentro de la va-
riable answer_sheets vista en el paso de la figura 14 poniendo el nombre en la primera posición.

```
answer_sheets = [
    'Formulario sin título (respuestas)',
    '16PF APQ (respuestas)',
    'EXPLORA (respuestas)',
    'IPP-R (respuestas)']
```

Figura 19. Asignación de respuestas nuevas

17. En las siguientes dos partes del código se debe cambiar el número que se encuentra allí por el 0.
    En este caso, se cambia el 2 por el 0, esto quiere decir que se tomará en cuenta el primer archivo que se encuentra en la variable answer_options_sheets

```
answer_options = options_extractor(answer_options_sheets[2])
```

Figura 20. Selección de opciones de respuestas

```
GetAnswers(
    answer_sheet=answer_sheets[2],
    answer_options=answer_options
```

Figura 21. Selección de respuestas

18. Se debe dirigir a la carpeta del proyecto, dar clic derecho, seleccionar Abrir en un terminal y escribir Python answers.py seguido de enter, con esto las respuestas de la hoja de respuestas creada en el paso 10 se extraen en un formato legible para el algoritmo.

19. Se abre el código de analysis.py y se cambia la variable k_clusters por el numero de grupos basados en la realidad. Es decir, si fueran 10 razas de perros, el número de k_clusters debería ser igual a 10, posteriormente se guarda y se ejecuta con el comando Python analysis.py, esto arroja el resultado de los clustroides para determinar una separación en los datos.

20. Si se quiere usar el método Elbow, se deben seguir las indicaciones del paso anterior con el archivo pruebas.py y ejecutarlo con Python pruebas.py, esto arrojará una gráfica, donde el ángulo más pronunciado de la curva indica una supuesta cantidad óptima de clusters.

4. METODOLOGÍA

Para el desarrollo de la metodología lo primero que se hizo fue tener en cuenta que se quería aplicar un modelo para análisis de datos, y que este fuese eficiente, se comenzó entonces con una base de datos pequeña de mil datos para probarlos en la etapa de análisis, a estos datos extraídos de una hoja de cálculo se les hizo el proceso de limpieza, modificando aquello que no se requería, estorbaba o necesitaba un cambio para el correcto análisis de los mismos. En este caso se separó la cabecera, es decir las preguntas de las respuestas, para usar las respuestas de cada individuo como un cluster que pudiera ser analizado. A parte también se hizo la transformación de datos de texto a numéricos pues los algoritmos de análisis funcionan mejor con datos numéricos, ya que se requieren tomar medidas de distancias. La limpieza de los datos es un paso muy importante, pues podría llegar a ocurrir que al estar mal estructurados, estos puedan llegar a ocasionar problemas en la ejecución de los algoritmos que usan dichos datos [35].

Se procedió entonces a realizar los algoritmos necesarios para el análisis de dichas respuestas, se tuvieron en cuenta varios aspectos, primero un enfoque basado a analizar datos que fuera eficiente al momento de procesar en términos de velocidad. Segundo, la dimensionalidad de los datos, por esto se procedió no solo a utilizar Clustering jerárquico y agrupamiento KMeans, sino que al momento de utilizar dichos algoritmos se tuvo en cuenta que dichos datos se encontraban en una dimensión que requerían medidas para espacios no euclidianos, por este motivo, se consideraron las distancias como Jaccard, Cosine, Hamming y Levenshtein. En el agrupamiento jerárquico se utilizó la diferencia hamming por simplicidad, una medida de distancia que permite hallar fácilmente diferencias entre datos de dimensión elevada, además de ser una distancia que funciona cuando los datos comparados son del mismo tamaño, por otro lado se hubiera preferido usar la distancia levenshtein, pues esta ultima permite hacer diferencias cuando los datos son de distinto tamaño.

El primer paso de la función principal, consiste en que esta llama a la función que se encarga de leer los datos limpios y traducidos previamente extraídos.

Después llama al método de agrupamiento jerárquico. Como se explicó en la sección 3.9 del marco teórico y conceptual y une los datos cuyas distancias sean mínimas, al unir estos datos a partir de las distancias, se garantiza que dichos datos sean lo mas semejantes posibles entre si. Es como si agrupáramos a los perros Golden Retriever porque las diferencias entre ellos son menores que si los comparásemos con perros Shih Tzu, entonces los Shih Tzu estarían agrupados en su propio cluster al determinar que sus distancias son menores entre si. Esta analogía se puede interpretar y utilizar también en este proceso, donde un cluster de personas determina un tipo de personas con algo en común. Este algo en común, representa la etiqueta que se le otorga a cada cluster después de hacer el análisis jerárquico. Este análisis también se puede entender como análisis no supervisado, pues no se conocen en principio que tipos de grupos existen en los datos que se deben procesar, es sino el usuario que al fin de este análisis les pone una etiqueta.

Para analizar se toman en cuenta todos los datos y se comparan utilizando la distancia Hamming, la idea es que se van agrupando los datos hasta tener un solo cluster, la distancia Hamming es la cantidad de modificaciones que requiere una serie de bits para ser igual a otra. Por ejemplo, la palabra 'paño' requiere de 1 modificación en la letra 'p' para ser igual a la palabra 'baño'. Esta distancia es utilizada para hallar la distancia inicialmente entre un par de datos, esta distancia está mejor explicada en la sección 3.6.10 del marco teórico y conceptual, esto quiere decir que a medida que toma distancias, se generan nuevos clusters. En el ejemplo tanto 'paño' como 'baño' pasarían a ser parte del mismo cluster al solo requerir de una modificación, mientras que por ejemplo 'bota' y 'gota' estarían en otro cluster.

Inicialmente todos los datos están separados, por lo que en teoría es como si cada dato fuese un cluster de un dato y al mismo tiempo su propio centroide, como cada dato es su propio centroide, inicialmente se hayan las distancias entre estos centroides. Recordemos que un centroide significa un punto que es la distancia media entre todos los puntos de un cluster, pero como solo hay un punto en el cluster, ese punto es el mismo centroide. Resultando algo similar a una tabla de distancias usando la métrica Hamming entre todos los puntos, como se muestra en la tabla 1.

Tabla 1. Distancias de Hamming.

distancias	p1	p2	p3	p4
p1	0			
p2	h(p1,p2)	0		
p3	h(p1,p3)	h(p2,p3)	0	
p4	h(p1,p4)	h(p2,p4)	h(p3,p4)	0

Luego se procede a usar el método completo para actualizar la matriz de distancias y unir clusters entre si, el método completo lo que hace es seleccionar la máxima distancia entre los miembros de un par de clusters, para ello primero seleccionamos el valor de la distancia hamming mínima y a esa combinación de puntos le aplicamos el método completo.

Supongase ahora que la distancia mínima en la tabla sea $h(p2, p3)$, estos dos puntos pasan a ser enlazados en un nuevo cluster. El siguiente paso sería recalcular la matriz de distancias, y para ello se toma los valores máximos de las distancias de $h(p2, p3)$ a los demás puntos del cluster, algo así $MAX(h(p2, p3), p1)$, $MAX(h(p2, p3), p4)$, $MAX(h(p2, p3), p5)$, $MAX(h(p2, p3), p6)$ y se actualiza como se muestra en la tabla 2.

Tabla 2. Método Completo.

distancias	p1	p2,p3	p3	p4
p1	0			
p2, p3	MAX(h(p2, p3), p1)	0		
p4	h(p1,p4)	MAX(h(p2, p3), p4)	h(p3,p4)	0

El siguiente paso sería tomar de nuevo la distancia Hamming mínima en la tabla y enlazar los puntos que tengan esa distancia para volver a recalcular con el método completo las nuevas distancias en la tabla, esto se repite hasta obtener un solo cluster con todos los puntos.

Esto nos permite tener acceso a una matriz de distancias, de las cuales escogeremos una para ser utilizada al momento de hallar un cluster representado en una dimensión, obtenemos este cluster pasándole a la función el resultado de aplicar Linkage como primer parámetro y la distancia escogida como segundo parámetro. Linkage es una función que agrupa los datos usando Clustering jerárquico. Esta función devuelve una matriz con las etiquetas de los clusters generados en cada paso y sus correspondientes distancias.

El Flat Cluster es el cluster resultante de una dimensión y es una matriz llena con las etiquetas que representan a cada uno de los clusters de la función Linkage, por lo tanto el tamaño de este es igual al número de datos que se están procesando. Recordemos que las etiquetas no son sino el nombre que se le otroga a cada cluster, en este caso las etiquetas se generan automáticamente y estas son numéricas.

Entonces, de este Flat Cluster escogemos una de las etiquetas de los clusters en representación entera y la usamos como el número de la cantidad de clusters que necesitamos obtener en KMeans, más adelante se mostrará como escoger el número de KMeans y la distancia en la sección de Experimentos y Resultados.

El segundo paso de la función principal, consiste en aplicar el método KMeans a los datos, para agruparlos en una cantidad k determinada, el algoritmo KMeans comienza escogiendo unos centroides en los datos de manera aleatoria, para casos prácticos en la sección de experimentos se inicializan los centroides.

El algoritmo KMeans como se explicó en la sección 3.10 del marco teórico y conceptual, se basa en los centroides para determinar el grupo al que pertenecen los datos y al igual que en el Clustering jerárquico los centroides iniciales son dos datos del cluster de datos, aunque también pueden ser datos en el universo posible, es decir datos imaginarios, la idea es medir la distancia euclidiana común 3.6.5 que hay entre cada dato en el universo y los centroides existentes, de esta manera se añade el dato al centroide cuya distancia haya sido mínima, por ejemplo, supongamos que de esta tabla de distancias euclidianas la distancia menor sea de $E_d(d3, c1)$, esto quiere decir que el dato 3 hará parte del cluster c1, como se muestra en la tabla 3.

Tabla 3. Distancia Euclidiana entre centroides.

datos/centroides	c1	c2
d1	$E_d(d1, c1)$	$E_d(d1, c2)$
d2	$E_d(d2, c1)$	$E_d(d1, c2)$
d3	$E_d(d3, c1)$	$E_d(d1, c3)$

El siguiente paso del algoritmo es recalcular los centroides siendo estos los centros de masa de

los datos de un cluster, los centroides se calculan tomando en cuenta los datos en matrices, por ejemplo, tomando una lista de listas de datos $data = d_1, d_2, ..., d_n$, donde cada dato tiene n datos $d_n = x_1, x_2, ..., x_n$, los centroides estarían representados en la misma forma de los datos $C_n = x_1, x_2, ..., x_n$, siendo la fórmula para calcular los centroides la siguiente $C_n = (x_1, ..., x_n)$ donde cada $x_n = \frac{\sum_{i=1}^{n} d_i}{n}$ para todo d_n en $data$.

De esta manera los datos se agrupan en clusters de características similares para su posterior análisis y etiquetamiento. Una vez obtenidos los clusters, hay una función que devuelve los clustroides a partir de los centroides, se trata de hallar la distancia Hamming entre los datos del cluster y el centroide, para escoger el dato mas cercano al centroide, este dato pasaría a ser el clustroide del cluster, lo que en ultimas sería el mismo proceso que para añadir un dato al cluster, solo que esta vez la distancias se toman con los datos existentes en los clusters y no con los que están afuera de ellos, esto quiere decir que el punto medio del cluster pasa a ser un posible punto imaginario a un punto real en el cluster, esto es bastante práctico en este modelo, donde se tomaría a un individuo como punto medio. Individuo que al ser etiquetado por el usuario que analiza los datos, también estaría etiquetando al grupo al que pertenece dicho individuo.

En el diagrama de la figura 22 se puede observar la fase de análisis.

El único proceso del usuario sería el de etiquetar los clusters generados y definir la cantidad de clusters que necesita, en el caso de análisis psicométricos se podría en cierta medida hacer caso omiso de los barémos. Los barémos son tablas que nos ayudan a determinar dependiendo de las respuestas de cada individuo a que grupo pertenece. Se puede entender que los baremos funcionan para individuos y no grupos de individuos, utilizando esta metodología se agrupan individuos con características similares, el trabajo de un psicólogo correspondería a realizar el baremo de los individuos que representan dichos clusters, para etiquetar a los demás individuos del cluster con el mismo baremo del individuo representativo, por ende se puede ahorrar el tener que utilizar un baremo para cada uno de los indiviuos por separado.

Otra técnica de análisis podría ser el coger cada individuo y hacer un baremo automático utilizando código de programación, pero este proceso sería increíblemente lento, pues el calculo

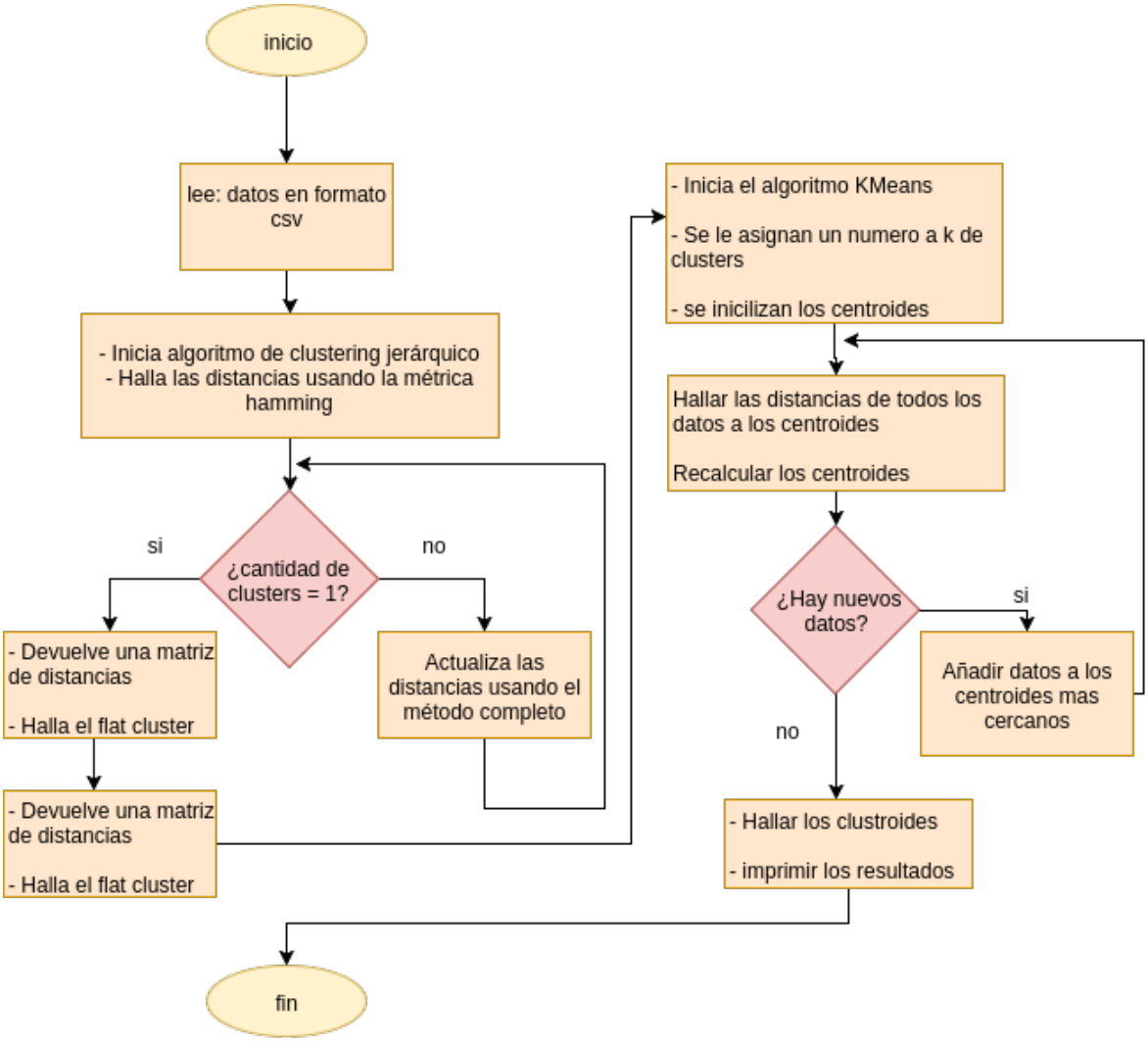

Figura 22. Fase de análisis

de cada baremo, representaría tiempo de computo innecesario.

5. EXPERIMENTOS Y RESULTADOS

5.1. Generando Datos

Para el desarrollo los experimentos se simularon los datos de las pruebas a realizar utilizando la librerías gspread, random, csv y oauth2client, así como la App de Google llamada Google Drive API que permite el acceso a las hojas de cálculo de google y todas sus características, de esta manera, se tomaron las posibles 180 respuestas de un set de preguntas sicométricas en formato csv y se generaron 1000 datos con 180 respuestas cada uno con un script realizado en Python, y posteriormente se subieron dichas respuestas a una hoja de cálculo en Google Sheets. Dicha hoja de respuestas simulada es la hoja que los usuarios como por ejemplo los psicólogos deberían tener previamente, esta puede ser generada por cientos de usuarios llenando una encuesta en Google Forms, pero para casos prácticos se simularon los datos, pues no se tenía el tiempo ni los recursos para que mil personas respondieran la encuesta.

Sea que la hoja de cálculo estuviera simulada o no, el siguiente paso fue crear la parte del código que arrastrara dichos datos y los convirtiera a un formato legible para los algoritmos a utilizar, por esta razón respuestas múltiples como si o no, se transformaron a 1 o 0, y así sucesivamente, hasta obtener un set de datos numéricos fácilmente manejable.

Se generaron pues, utilizando como base las pruebas psicométricas IPP-R, EXPLORA y 16PF-APQ de las cuáles se habló en el marco teórico conceptual, 3 set de mil datos que contenían de 180 a 200 respuestas cada uno.

En resumen se generó un script para crear respuestas aleatorias, cargarlas en una base de datos en Google Sheets y extraerlas en un formato legible para los procedimientos necesarios, un ejemplo sobre como acceder a los datos de una hoja de calculo de Google se encuentra en la sección de Google Spreadsheets Python API en el marco teórico y conceptual.

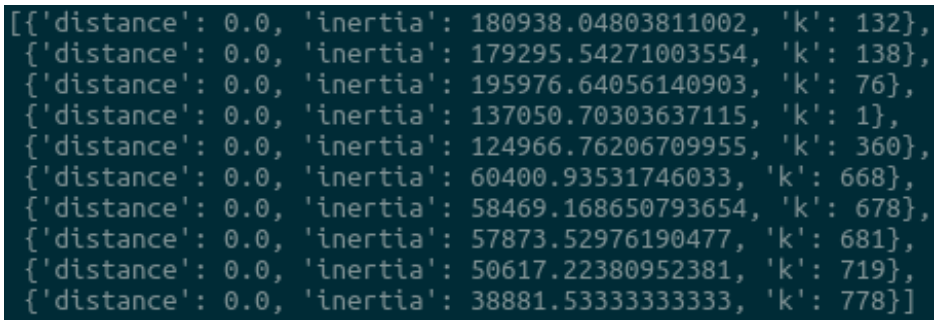

Figura 23. Valor optimo de distancia

5.2. Aplicando Algoritmos

Una vez los resultados de los test psicométricos se encontraron correctamente traducidos a datos numéricos, se procedió a modificar los centroides, pues en el algoritmo KMeans fue necesario inicializar los centroides en vez de que estos fueran seleccionados arbitrariamente, pues como los datos tienen medidas tan semejantes, era un problema al momento de hacer las pruebas y arrojar resultados con distintas cantidades de clusters, la razón es que aveces se unían unos datos a ciertos clusters y otras veces a otros clusters, lo que imposibilitaba saber con determinación en varias iteraciones si los clusters eran adecuados para un análisis visual de los mismos.

Utilizando las distancias otorgadas por la función Linkage se procedió a analizar que medida de distancia provocaba una función de costo más reducida iterando las distancias sobre los clusters dados y evaluando el valor de la inercia que fuera mínima, dando como resultado que el valor óptimo siempre tendía a cero, por lo que se procedió a utilizar esa constante.

Se tomó como criterio entregar 3 veces el número de clusters pedido para que el usuario pueda separar los datos y encontrar pequeñas semejanzas, que le permitan hacer observaciones, generar nuevos grupos y juntar datos entre si, por esta razón se entrega el resultado de los clustroides de cada cluster, pues los clustroides a diferencia de los centroides, son valores reales en el cluster, es así como se obtiene un individuo real promedio de la muestra de datos y clasificando a dicho individuo, se clasifica también el cluster al que pertenece dicho individuo.

Los clustroides se obtuvieron tomando la distancia mínima de las distancias de todos los puntos

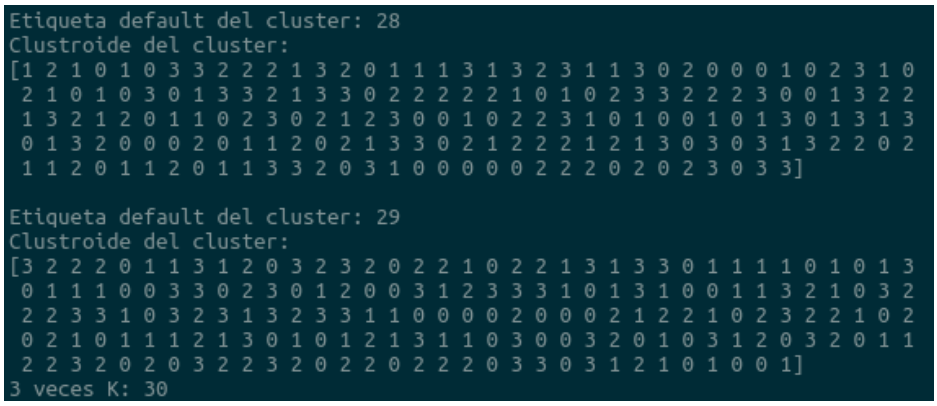

Figura 24. Clustroides

de cada cluster a su centroide, la distancia utilizada para hacerlo fue la distancia Hamming. En resumen, los datos del clustroide son respuestas de un individuo común que representa a los demás individuos del grupo.

5.3. Pruebas

Para el desarrollo de las pruebas se utilizó el código generado para los algoritmos y se iteró sobre el mismo, ensayando toda la cantidad de clusters posibles en el rango de la cantidad de datos, tratando de encontrar de manera visual la función codo (THE ELBOW), pues encontrando un codo nos indicaría hasta donde los clusters se reparten de manera sustancial, después del codo los clusters se convierten en ruido indica una separación incorrecta de los datos. La linea azul, indica la cantidad de clusters seleccionados por el observador en el algoritmo. La curva más pronunciada en cualquiera de las gráficas, si se encuentra uno, indica hasta que punto podríamos aumentar el número de clusters y tener una separación de datos de acuerdo a las características.

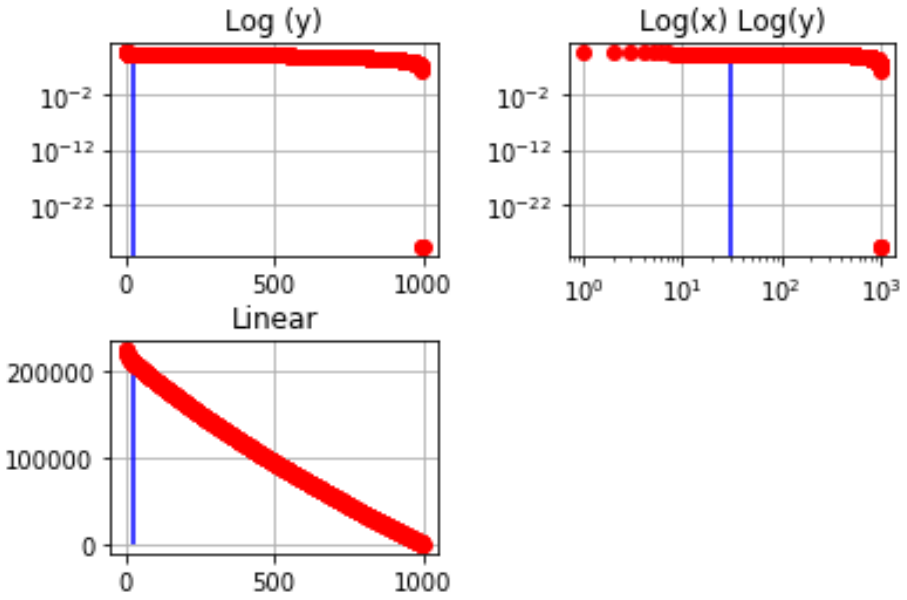

Figura 25. Prueba del codo

6. CONCLUSIONES Y RECOMENDACIONES

6.1. CONCLUSIONES

- Los métodos utilizados para el análisis de los datos funcionan muy bien cuando los datos caben en memoria.

- Cuando los datos no caben en memoria se deben utilizar otros métodos como el algoritmo *THE CURE* explicado en el marco teórico conceptual, ya que reduce el tamaño de la muestra al procesar solo los puntos representativos. Los métodos utilizados en este documento hacen parte del algoritmo *THE CURE*.

- Al momento de hallar los clustroides el algoritmo *THE CURE* sería más rápido puesto que no se usan todos los datos para hallar la distancia al centroide, sino que se usan solo los datos representativos de cada cluster.

- Se pueden utilizar diferentes algoritmos para calcular las distancias, pero la distancia Hamming es perfecta cuando se trata de medir distancias entre datos de igual tamaño, si los datos fuesen de tamaño diferente se podría utilizar la distancia Levenshtein.

- El método *Elbow* no siempre produce un codo que se acomode y muestre la cantidad de clusters ideal, pues aveces dichos datos producen resultados muy secuenciales que no muestran un cambio sustancial en la inercia.

- La alta dimensionalidad de los datos y el criterio de las respuestas inhibe que se escojan datos para una representación bidimensional de los cluster en una gráfica.

- Este enfoque basado en características recogidas pertenecientes a individuos, anonimiza la muestra de los datos, haciendo que la metodología aplicada se pueda utilizar en otras áreas como la medicina, el transporte, la política, entre otros.

- Una base de datos puede ser considerada *big data* cuando no cabe en la memoria del procesamiento del computador que la corra, pero un algoritmo de *big data* puede correrse en

una base de datos que si quepa en memoria, en cuyo caso la velocidad del procesamiento de los mismos podría verse considerablemente afectada.

6.2. RECOMENDACIONES

- Para el desarrollo de los algoritmos necesarios, es recomendable revisar primero si las librerías de Python que ya implementan ciertas funcionalidades.
- Se recomienda tener un entendimiento general de las bases de datos que se analizan para aplicar los algoritmos de manera correcta, una mala interpretación de los datos daría un modelo que no se ajusta a la realidad.
- El entendimiento sobre que distancias aplicar de acuerdo a los datos que se tienen es muy importante a la hora de determinar como proceder con los algoritmos.
- Spyder es un entorno de desarrollo de Python científico que facilita no solo la visualización de los datos sino también la limpieza del código.

BIBLIOGRAFÍA

[1] MICHELL, Joel. Measurement in Psychology: A Critical History of a Methodological Concept. Ideas in Context. Cambridge University Press, 1999. 1

[2] BIJSMANS, E.S., *et al.* Psychometric Validation of a General Health Quality of Life Tool for Cats Used to Compare Healthy Cats and Cats with Chronic Kidney Disease. En: Journal of Veterinary Internal Medicine, tomo 30, nº 1, 2016, págs. 183–191. ISSN 1939-1676. 1

[3] ON STANDARDS FOR EDUCATIONAL EVALUATION, Joint Committee; GULLICKSON, A.R. y OF SCHOOL ADMINISTRATORS, American Association. The Student Evaluation Standards: How to Improve Evaluations of Students. 1-Off Series. SAGE Publications, 2003. ISBN 9780761946632. 1

[4] IMMEKUS, J. C. y FRENCH, B. F. Psychometrics in engineering education: evaluating test score reliability and validity. En: 34th Annual Frontiers in Education, 2004. FIE 2004., 2004. ISSN 0190-5848, págs. 1465–1465. 1

[5] LANGLEY, Pat. The changing science of machine learning. En: Machine Learning, tomo 82, nº 3, 2011, págs. 275–279. ISSN 1573-0565. 1

[6] SUTTON, Richard S. y BARTO, Andrew G. Introduction to Reinforcement Learning. 1ª ed$^{\text{ón}}$. MIT Press, Cambridge, MA, USA, 1998. ISBN 0262193981. 1

[7] BISHOP, Christopher M. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738. 1

[8] HINTON, G., *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. En: IEEE Signal Processing Magazine, tomo 29, nº 6, 2012, págs. 82–97. ISSN 1053-5888. 1

[9] BOTTOU, L., *et al.* Comparison of classifier methods: a case study in handwritten digit recognition. En: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5), tomo 2, 1994, págs. 77–82 vol.2. 1

[10] WARD, Jonathan Stuart y BARKER, Adam. Undefined By Data: A Survey of Big Data Definitions. En: CoRR, tomo abs/1309.5821, 2013. 1

[11] LOHR, Steve. The Age of Big Data, 2012. [Online; posted 11-February-2012]. 1.1

[12] BURN-MURDOCH, John. Is the recruitment industry set for a big data revolution?, 2013. [Online; posted 8-August-2013]. 1.1

[13] LEMUS, Mauricio Vega. Pereira en la Era del Big Data, 2017. [Online; posted 28-August-2017]. 1.1

[14] BIO, Full. The Complete Beginner's Guide To Big Data In 2017, 2017. [Online; posted 14-March-2017]. 1.1

[15] DE COLOMBIA, Universidad Nacional. Informe de Gestión 2016. En: Gestion, tomo 1, 2016, pág. 139. 1.1

[16] DE RISARALDA, Corporación Autónoma Regional. Belen de Umbria: Datos Generales, 2017. [Online; posted 04-September-2017]. 1.1

[17] EPS, Nueva. Redición de Cuentas 2016, 2017. [Online; posted April-2017]. 1.1

[18] MARKOWETZ, Alexander, *et al.* Psycho-Informatics: Big Data shaping modern psychometrics. En: , tomo 82, 2014. 2

[19] HERLAND, Matthew; KHOSHGOFTAAR, Taghi M. y WALD, Randall. A review of data mining using big data in health informatics. En: Journal Of Big Data, tomo 1, nº 1, 2014, pág. 2. ISSN 2196-1115. 2

[20] CHEUNG, Mike W.-L. y JAK, Suzanne. Analyzing Big Data in Psychology: A Split/Analyze/Meta-Analyze Approach. En: Frontiers in Psychology, tomo 7, 2016, pág. 738. ISSN 1664-1078. 2

[21] CHEN, Eric y WOJCIK, Sean. A Practical Guide to Big Data Research in Psychology. En: , tomo 21, 2016, págs. 458–474. 2

[22] MEHRENS, William A y LEHMANN, Irvin J. Using standardized tests in education, 4th ed. Longman/Addison Wesley Longman, New York, NY, US, 1987. ISBN 0-582-29022-8 (Hardcover), xi, 529–xi, 529 págs. 3.2, 3.3

[23] M, Martínez Vicente J. EXPLORA CUESTIONARIO PARA LA ORIENTACIÓN VOCACIONAL Y PROFESIONAL. International Journal of Developmental and Educational Psychology. URL `http://www.redalyc.org/articulo.oa?id=349851787037`. 3.4.1

[24] Y DAVID ARRIBAS ÁGUILA, Fernando Sánchez Sánchez. BAT-7, BATERÍA DE APTITUDES DE TEA: DESCRIPCIÓN Y DATOS PSICOMÉTRICOS. En: International Journal of Developmental and Educational Psychology Revista INFAD de Psicología, tomo 2, nº 1, 2016, págs. 353–364. 3.4.2

[25] FERRER, Laia y KIRCHNER, Teresa. Suicidal Tendency Among Adolescents With Adjustment Disorder. En: Crisis, tomo 36, nº 3, 2015, págs. 202–210. 3.4.3

[26] CIPRIANO, O.; PATRICIA, U. y NADIA, D. Persona. En: 16, 2013, págs. 139–164. 3.4.4

[27] GOWER, J.C. Properties of Euclidean and non-Euclidean distance matrices. En: Linear Algebra and its Applications, tomo 67, nº Supplement C, 1985, págs. 81 – 97. ISSN 0024-3795. 3.5

[28] WU, X., *et al.* Data mining with big data. En: IEEE Transactions on Knowledge and Data Engineering, tomo 26, nº 1, 2014, págs. 97–107. ISSN 1041-4347. 3.6

[29] GANTI, V., *et al.* Clustering large datasets in arbitrary metric spaces. En: Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337), 1999. ISSN 1063-6382, págs. 502–511. 3.6.1, 3.6.2

[30] BOLSHAKOVA, N. y AZUAJE, F. Cluster validation techniques for genome expression data. En: Signal Processing, tomo 83, nº 4, 2003, págs. 825 – 833. ISSN 0165-1684. 3.6.3, 3.6.4

[31] LJUBESIC, N., *et al.* Comparing measures of semantic similarity. En: ITI 2008 - 30th International Conference on Information Technology Interfaces, 2008. ISSN 1330-1012, págs. 675–682. 3.6.5, 3.6.6, 3.6.7, 3.6.8, 3.6.9, 3.6.10, 3.6.11

[32] WALLER, Matthew A. y FAWCETT, Stanley E. Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. En: Journal of Business Logistics, tomo 34, nº 2, 2013, págs. 77–84. ISSN 2158-1592. 3.7, 3.8

[33] ROSE, Kenneth; GUREWITZ, Eitan y FOX, Geoffrey C. Statistical mechanics and phase transitions in clustering. En: Phys Rev Lett, tomo 65, 1990, págs. 945–948. 3.9

[34] FAHIM, *et al.* An efficient enhanced k-means clustering algorithm. En: Journal of Zhejiang University-SCIENCE A, tomo 7, nº 10, 2006, págs. 1626–1633. ISSN 1862-1775. 3.10, 3.11

[35] JAGADISH, H. V., *et al.* Big Data and Its Technical Challenges. En: Commun ACM, tomo 57, nº 7, 2014, págs. 86–94. ISSN 0001-0782. 4